DOCUMENT RESUME

ED 156 110 IR 005 824

AUTHOR	Gaus, Wilhelm, Ed.; Henzler, Rolf, Ed.
TITLE	Data Documentation: Some Principles and Applications in Science and Industry. Proceedings of the Workshop on Data Documentation Organized by the School for Medical Documentation of the University of Ulm.
SPONS AGENCY	Commission des Communautes Europeennes (Luxembourg).
PUB DATE	77
NOTE	187p.
EDRS PRICE	MF-$0.83 HC-$10.03 Plus Postage.
DESCRIPTORS	Chemistry; *Computer Oriented Programs; Confidentiality; *Data; *Data Bases; *Documentation; Electronic Data Processing; Foreign Countries; *Industry; Information Retrieval; Information Storage; Programing Languages; *Statistical Analysis

ABSTRACT

The 13 papers collected here were presented at the Workshop on Data Documentation held at Reisenburg Castle, Federal Republic of Germany, in July 1975. This workshop was a first attempt to identify some of the problems in the field of data documentation, formulate them, and solve them as far as possible. The subject was approached from four different viewpoints: (1) from the point of view of a documentalist, and from documentation theory; (2) the EDP as the most important tool with a strong influence on the documentational concept itself; (3) management and organizational aspects; and (4) presentation of existing data documentation systems in such areas as chemistry, medicine, steel and iron materials, and personal data. Additional special topics addressed included statistical methods for data documentation and confidentiality of personal data. These papers were written for readers with advanced knowledge in at least one of the subjects touched on by data documentation. The proceedings conclude with a bibliography based on the citations of the authors, a bibliography from the German Society for Medical Documentation and Statistics, and searches from the Central Library for Documentation of the German Society of Documentation and some smaller sources. (Author/JPF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED156110

Commission of the European Communities

Data Documentation: Some Principles and Applications in Science and Industry

Proceedings of the Workshop
on Data Documentation organized
by the School for Medical Documentation
of the University of Ulm,
(Federal Republic of Germany)
Reisensburg Castle, July 7–11, 1975

Edited by
Wilhelm Gaus and Rolf Henzler

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Wilhelm Gaus

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

IR005824

VD
Verlag Dokumentation, Publishers, München 1977

EUR 5656e

The Commission of the European Communities,
Directorate-General Scientific and Technical
Information and Information Management,
Luxembourg

LEGAL NOTICE
Neither the Commission of the European Communities
nor any person acting on behalf of the Commission is
responsible for the use which might be made of the
following information.

Editors
Wilhelm Gaus, University of Ulm
Rolf Henzler, Deutsches Krebsforschungszentrum, Heidelberg

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Data documentation : some principles and applications in science and industry ; proceedings of the Workshop on Data Documentation organized by the School for Med. Documentation of the Univ. of Ulm (Federal Republic of Germany) Reisensburg Castle, Juli 7 – 11, 1975 / Comm. of the Europ. Communities. Ed. by Wilhelm Gaus and Rolf Henzler. – 1. Aufl. – München : Verlag Dokumentation, 1977.
ISBN 3-7940-3022-2

NE: Gaus, Wilhelm [Hrsg.]; Workshop on Data Documentation ⟨1975, Reisensburg⟩; Schule für Medizinische Dokumentations-Assistenten ⟨Ulm⟩; Europäische Gemeinschaften / Kommission

© ECSC, EEC, EAEC Luxembourg, 1976
Publisher: Verlag Dokumentation Saur KG, München
Printed by Hain-Druck KG, Meisenheim/Glan
Bound by Thomas-Buchbinderei GmbH, Augsburg
Federal Republic of Germany
ISBN 3-7940-3022-2

Acknowledgements

The Workshop on Data Documentation was sponsored by the Commission of the European Communities, Directorate-General Scientific and Technical Information and Information Management, and by the Institute for Documentation (Frankfurt). The Commission of the European Communities also assisted in the publication of this volume. The International Institute for Scientific Cooperation, Reisensburg Castle, was the host of the meeting and contributed to it with a stimulating atmosphere. We offer our sincere thanks to all our sponsors.

We also thank the lecturers for presenting their papers, the staff of the School for Medical Documentation for organizing the workshop at rather short notice and, last but not least, we thank the participants and lecturers for their involvement in the subject. This involvement expressed itself in several informal discussions which often continued long after midnight.

The Editors

Contents

Appleyard, Raymond K.:
Foreword 8

Editorial 9

Andersen, Per Klitgaard:
Data Retrieval – from the Management's Point of View 13

Dubois, Jacques-Emile; Jean-Claude Bonnet; Jean-François Guaydier:
The DARC System: A Methodology for Data Bank Organization . . 23

Dathe, Gerd; Karl-Heinz Dreckmann:
MAINS Iron & Steel: An Information System for Iron and Steel Materials 25

Gaus, Wilhelm:
Statistical Methods in Data-Documentation 48

Killian, Klaus-Peter:
Clinical Laboratory Data, Capture and Evaluation 53

Konrad, Erhard:
Principles of Information Retrieval Systems 60

Lockemann, Peter C.:
Retrieval Languages 64

Lutz, Theo:
Structures of Computerized Data Bases 66

Schäfer, Christoph:
Data Documentation – an Overview 97

Schlörer, Jan:
Confidentiality and Security in Statistical Data Banks 101

Schmidt, Herbert:
Social-Information-System in the Federal Republic of Germany . . 124

Weingarten, Werner:
Hospital Information System Hanover 145

Wersig, Gernot:
Contributions of Information Science to Data Documentation 155

Bibliography 162

List of Authors 188

Foreword

Information, that is the result of the research undertaken by either its scientists or by others, has become one of Europe's most precious resources. Our difficulty and also our challenge is that information science on techniques for handling this precious commodity is still in its infancy and we are a long way from being able to make efficient use of it.

The Member States of the European Communities are well aware of this problem and also of the obvious means of attacking it, one of which is to provide more and better training of information specialists. The rôle of the Community in these matters is above all one of coordination to make sure that the best match possible is available within the Nine countries between training courses and students seeking instruction.

By choosing to support the workshop on data documentation organized by the School for Medical Documentation at the University of Ulm, the Commission recognized that data documentation, although not very widely established in Europe, is emerging as an important part of information technology; clearly, this experiment was a concrete instance of low cost and effective sharing of progress.

The truly international response of the specialists who attended the course either to teach or to learn is gratifying both to its organizers and to its sponsors.

Raymond K. Appleyard, Luxembourg

Editorial

Education and training are the most important investments for the future. In 1973 the Commission of the European Communities established within the Committee for Information and Documentation in Science and Technology (CIDST) the "Working Group Training of Specialists", to promote information science in its member states. The main goals of this working group are to improve the knowledge of information and documentation science, to harmonize the European scientists in this field, and to prepare them for closer knit cooperation. For this reason the working group supported the following workshops in 1975:
Educational Methods (United Kingdom)
Data Documentation (Germany)
Multilingual Thesauri (France)
Linguistic Aspects (Denmark)

The "Workshop on Data Documentation" was a first attempt to identify the wide spread problems in the field of data documentation, formulate them, and to solve them as far as possible. It was organized by the School for Medical Documentation at the University of Ulm and sponsored by the Commission of the European Communities and the Institute for Documentation in Frankfurt. In July 1975 16 delegates and 14 lecturers met at Reisensburg Castle, the guesthouse of the International Institute of Scientific Cooperation, Bavaria, Germany, for a one week workshop (July 6 to July 14).

The subject was approached from four different viewpoints:

- from the point of view of a documentalist, and from documentation theory
- the EDP as the most important tool with a strong influence on the documentational concept itself
- management and organizational aspects
- presentation of existing data documentation systems in chemistry, medicine, steel- and iron materials, personal data etc.

Additional special topics were also on the agenda, such as

- statistical methods for data documentation
- confidentiality of personal data.

Of course we were unable to solve all problems faced. Especially straight forward discussion between documentalists and computer scientists is not always easy to achieve. But there is no doubt that data documentation can only be performed well by the cooperation of several specialists, e.g. documentalists, computer scientists, programmers, statisticians, managers and especially professionals of the field the data are coming from (or used for).

Data documentation is already very important for most of the work in industry, business, administration, health service, government, science, and research; but it will become more important in the next decade. One reason for the increasing importance of data documentation is an economic one: Small dedicated computer systems will reduce costs of the collection of machine readable data: increasing personal costs and decreasing computing costs will move the steady state between manual and machine work towards the computer; and it will become uneconomically expensive to make decisions without access to the appropriate data banks.

Data documentation is between the usual subjects and professions as already mentioned. The reader should not expect a complete and comprehensive presentation of the methods, techniques, possibilities, and usage of data documentation in this presentation. We were only able to do some parts in the complex and detailed picture of data documentation, hopefully some important ones and some of common interest. Therefore the following papers are written for readers with advanced knowledge in at least one of the subjects touched on by data documentation. For readers who want to introduce themselves to data documentation with this volume the following sequence in reading the contributions is recommended: Schäfer, Wersig, Weingarten, Kilian, Andersen, Schmidt, Schlörer, Gaus, Dathe & Dreckmann, Dubois et al, Lutz, Lockemann, Konrad.

Dr. *Andersen* (Denmark) who is the manager of the data processing department in a large banking house, deals in his contribution not only with the importance of good management, but also about its implementation. This is demonstrated with an example concerning the marketing of food. There is an influence in both directions: management of data documentation and data documentation for the management.

Mr. *Dathe* and Mr. *Dreckmann* (Germany) are engaged in an organisation of the German steel & iron industry, which provides its member firms with mechanical and chemical data of materials. They explain the computerized data documentation system and the use of this system in detail and give an outline of further developments planned.

Prof. *Dubois*, Dr. *Bonnet*, and Dr. *Guaydier* (France) from the Laboratoire de Chimie Organique Physique, Faculty of Science of the University of Paris describe a computerized system for acquisition, description and retrieval of data on chemical structures, developed in their institute during several years.

Dr. *Gaus* (Germany), head of the School for Medical Documentation and the Department of Medical Statistics, Documentation and Data Processing, both at the University of Ulm, deals with the most important statistical procedures in data documentation. Obviously it is often necessary to compress a lot of individual data into a generalized statement or figure. Further, statistical methods are used to determine correlations, dependencies, clusters and other immanent structures in the data.

Dr. *Kilian* (Germany) works in the chemical laboratory of a large new hospital in Munich. He shows us a highly automated and sophisticated data acquisition system, which picks up most of the data on line and in real time. Several checks are used to get correct data as far as possible. The collected data are stored, reorganized, and then presented in various ways.

Dr. *Konrad* (Germany) is from the Technical University of Berlin and points out the general principles of information retrieval with special emphasis of retrieval systems for unformated information. The concepts of documentation as used in document retrieval are matched with data base technology and modern techniques used in linguistic data processing.

Prof. *Lockemann* (Germany) from the faculty of computer science at the University of Karlsruhe talks in his paper about special programming and retrieval languages for the establishment, maintenance, and use of databases. The database is regarded as a model of the real world. The retrieval language is built up as a specific enlargement of a host language.

Mr. *Lutz* (Germany) employed by IBM, German headquarter at Sindelfingen, gives a detailed layout of computerized data banks for formatted data, their basic concepts, elements, file organisation, access, etc. He also touches on the enlargement of a programming language for data bank communication (host language) as well as a high level language for handling data banks (query language).

Dr. *Schäfer* (Germany), chairman of the Informationszentrum für Biologie, Forschungsinstitut Senckenberg at Frankfurt gives an introduction to data documentation, enumerates various definitions and categories of data and data documentation. It depends strongly on the use of the data and the type of data what the data documentation looks like. This is valid not only from the documentalist's point of view, but also for data acquisition, processing and presentation.

Dr. *Schlörer* (Germany), Department of Medical Statistics, Documentation, and Data Processing at the University of Ulm, is engaged in the confidentiality of anonymous personal data (data about individual persons but without name, address, etc.). He found possibilities to identify and therefore to de-anonymize those data only by asking the statistical data bank the usual types of queries. Only a little amount of preknowledge about the person of interest is necessary to identify the data of this person in a rather large data bank. Then all information stored about this person can be retrieved, although countermeasures might be set up.

Dr. *Schmidt* (Germany), Regierungsdirektor in the Ministery of Labour and Social Affairs, stresses the importance of good detailed information for the government. Therefore the Ministery of Labour and Social Affairs keeps a large joint social data bank with data about health insurance, accident insurance, social security insurance, employment services, etc. This social data bank is used not only for retrieval, but also for statistical evaluations and simulations of the socio-political decision-making process.

Mr. *Weingarten* (Germany), member of the Department of Medical Informatics at the Medical University of Hannover, describes the hospital information system developed and used by the Medical University of Hannover. The system is spread over most parts of the hospital and is operated by a large central computer which keeps the medical records of the patients in a data base. The detailed services of this hospital information system influence the organisation, the management, the administration, and of course the health care provided by the hospital.

Prof. *Wersig* (Germany), member of the Free University of Berlin, gives an overview of several approaches from information science to data documentation: information as structure, knowledge, message, understood message, etc. He also deals with similarities and differences between literature documentation and data documentation and how far the documentary languages used in literature documentation (e.g. classification, keywords, thesauri) are helpful in data documentation.

Due to the fact that we could deal in this volume with only a few aspects of data documentation we add a bibliography. The bibliography is based on the citations of the authors, a bibliography of the Deutsche Gesellschaft für Medizinische Dokumentation und Statistik (German society for medical documentationtion and statistics), and searches at the Zentralstelle für Dokumentation der Deutschen Gesellschaft für Dokumentation (Central Library for Documentation of the German Society of Documentation) at Frankfurt, and some smaller sources.

Data Retrieval – from the Management's Point of View

Per Klitgaard Andersen, Kobenhavn

Introduction

As an introduction to this paper about the managerial problems connected with the construction of data retrieval systems, I want to give my definition of a data retrieval system (Fig. 1). This figure shows, that I define a data retrieval system as a system which makes a qualitative processing of complicated nature of the data registered in the system. This should be seen in comparison with an information processing system, which makes a quantitative and uncomplicated processing, and an information retrieval system, which makes a quantitative and complicated processing.

The managerial problems with the construction of data retrieval systems are, that most organizations starting this work had earlier implemented systems of other types – data processing systems, information processing systems or information retrieval systems. Therefore the management has to pay special attention to the strategy (management of objectives), the organization (management of staff), and the standards (management of techniques) (Fig. 2). The experience shows, that these management problems are much more complicated when the development efforts of the organization systems are no longer working on virgin soil.

I should like to discuss these problems by mentioning four points: firstly an example of a data retrieval system, secondly a definition of a systems development strategy, thirdly a description of the organizational frameworks, and finally a discussion about the standards, which should be regarded both as construction tools and constraints in the system develoment work (Fig. 3).

Application example

The following example of a data retrieval system is taken from a danish industrial firm of medium size, which produces freeze-dried food products. The reason for choosing this example is that the system from a conceptual point of view is well suited to describe the managerial problems encountered in the construction of data retrieval systems: The system itself is not large, but the objectives and constraints for the system are typical.

The basic manufacturing principle for freeze drying (Fig. 4) is examplified with the treatment of mushrooms. The example shows, that the raw product is pre-

Data processing	qual – uncomplete
Information processing	quant – uncomplete
Data retrieval	qual – complete
Information retrieval	quant – complete

Fig. 1: Definitions and presuppositions

Management of objectives – Strategy
Management of staff – Organization
Management of techniques – Standards

Fig. 2: Management problems

1 Application example
2 Development strategy
3 Organization and staff
4 Standards

Fig. 3: Subjects

treated and deep-frozen, after which it is dehydrated in a vacuum cabinet. Quite obviously the final product contains only the dry-substance of the raw-material of the manufacturing process (Fig. 5): As the weight of the final product only amounts to 5–10% of the raw-material weight and the raw-materials are perishable, the whole manufacturing procedure depends very heavily on the raw-materials, both regarding quality, quantity, and delivery time. The objectives of the data retrieval system are therefore to create a planning tool for the managers of the individual departments (purchase, production, and sales), and a coordination tool for the top-management. The construction of the system has, of course, some constraints, mainly concerning that data should be easily acquired, and that the final system should fit into the administrative procedures of the firm.

The final data retrieval system was established on-line on a real time basis with the functions and specification indicated above (Fig. 6). Conceptually the system is a typical data retrieval system, because the treatment of the data is qualitative (computations are made, trends and plan deviations are calculated) and complicated (the structure of the data bank allows interrelation between all possible raw-materials and final products).

If you regard the above mentioned system as an isolated system, it is relatively easy to get an overall impression of the objectives of the system, and the demand satisfaction it will provide.

If the system is established in a portfolio of other administrative systems, this evaluation grows far more complicated, and that is exactly what happens to most data retrieval systems. In these cases it is essential to formulate a systems development strategy.

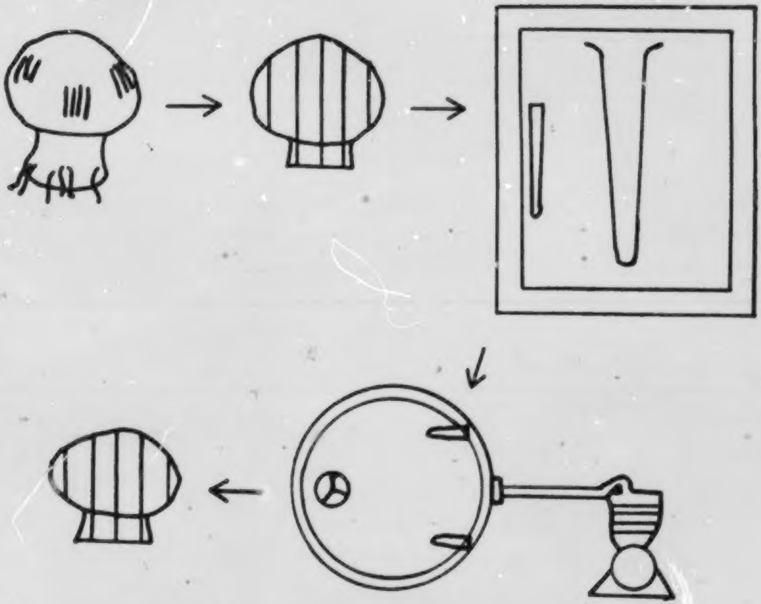


Fig. 4: Basic manufacturing principle

Raw materials
  Quality
  Quantitiy
  Delivery

Data retrieval
  Management
  Functional coordination

Constraints
  Data capture
  Administration procedures

Fig. 5: Problems and objectives

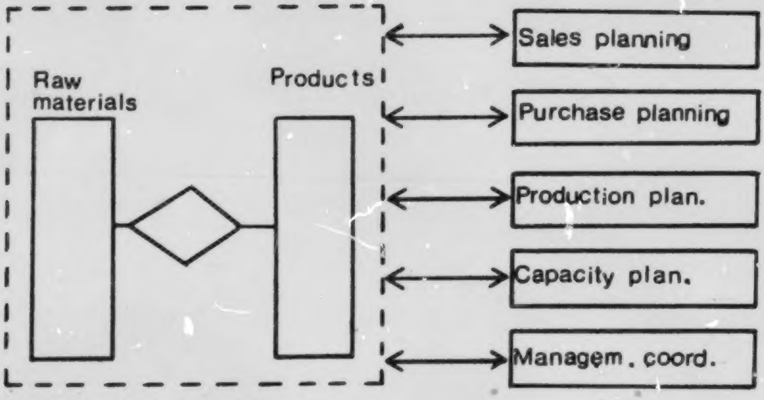


Fig. 6: Data base and functions

Systems Development Strategy

The purpose of a systems development strategy is to ensure that a new administrative system has objectives which both fit to the objectives of the organization and correspond to the other administrative systems already established in the organization. A systems development strategy consists of a model, describing the means and objectives of the different system categories, and some constraints, which shall secure the described objectives to be obtained by the system.

I want to emphasize that the formulation of a systems development strategy is a managerial job, and that no theory guarantees that it is possible to formulate it. When you want to include a new system among the others in your organization there are two possibilities: the new system either fits into your systems development strategy, or it does not. If the new system does not fit, you have to formulate a new strategy, and if this is not possible, you are admitting that it is not possible to formulate a systematization of the objectives for young administrative systems.

The model of a system can be formulated in many ways, and I want to describe an example, showing in which way it could be done in a private firm.

The firm in the example has 5 administrative systems (Fig. 7), and wants to introduce a new system, for example to support sales-promotion. After the means and objectives of the administrative systems have been described in a

Production planning and purchase	T
Invoice	O
Staff salary and tax	O
Profit and loss-budget	T
Consumer model	S

Fig. 7: System levels in manufacturing

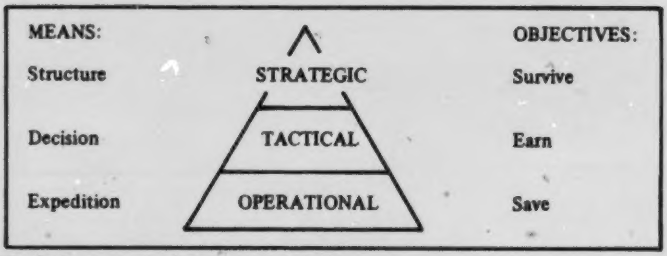


Fig. 8: Framework of development strategy

model (Fig. 8), the systems of the firm are divided in three main categories: operational systems, tactical systems, and strategic systems. In this case the new sales-promotion system is regarded as a tactical system which by means of providing a better decision basis had the objective of creating better incomes four your organization.

It is essential to understand that the systems development strategy is an important communication tool between management and system constructors. In the above mentioned case it is essential that both parts agree that the new system is certainly not planned to create savings by means of improvements in the expedition routines, and that they prepare their working dispositions according to these facts.

Finally a few words about the constraints which are a part of the system development strategy: these constraints concern the integration between the different system categories, the system technology in the different system categories (i.e. on-line or batch) and the EDP-technology in the different system categories (i.e. hardware and software manufacturer). These constraints should be formulated explicitly to create a conceptual basis in the communication between system constructors and management (Fig. 9 and Fig. 10).

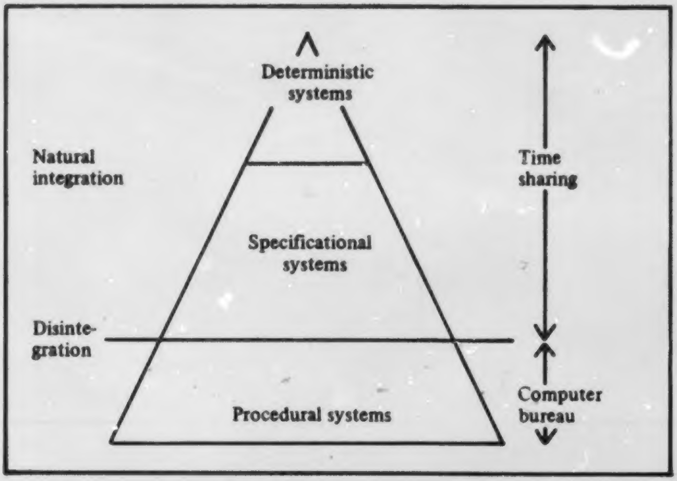

Deterministic systems
Natural integration
Time sharing
Specificational systems
Disintegration
Procedural systems
Computer bureau

Fig. 9: Development strategy – manufacturing company

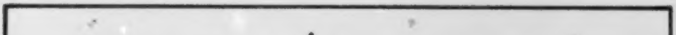

Deterministic systems
Time sharing
Disintegration
Specificational systems
Systematic integration
Own computer
Procedural systems

Fig. 10: Development strategy – commercial bank

Organization

As earlier mentioned data retrieval systems are seldom built on virgin soil, and this means that it is quite normal to have an organization which is experienced in development and use of other EDP-systems. The introduction of a new data retrieval system in the organization means new requirements, and as also staff changes will be going on usually, it is necessary that the management evaluates and plans the development of the organization. The purpose is to plan the organization for the next 5 years, but it is essential for the management to realize, that this plan cannot be based upon the reqirements of the system only, but of equal importance has to consider the past and the future of the staff, who will fulfil the requirements.

In the following I will discuss this organizatorial development by giving a brief status of the organization in 1965, 1970, and 1975 for a firm with its own EDP-department (Fig. 11).

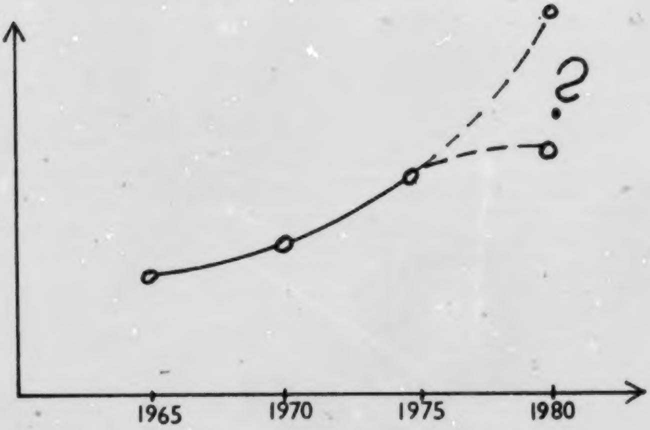


Fig. 11: Organization, changes in requirements, changes in staff, changes in organization

First of all it is essential to realize that most internal EDP-departments have been created not by EDP-specialists but by men originating within the firm and with a deep and practical understanding of the firm's administrative procedures. In 1965 such an EDP-department was typically organized in an operations, a programming department and a planning department, and as the staff knew the administrative procedures of the firm, the organization normally performed quite well (Fig. 12).

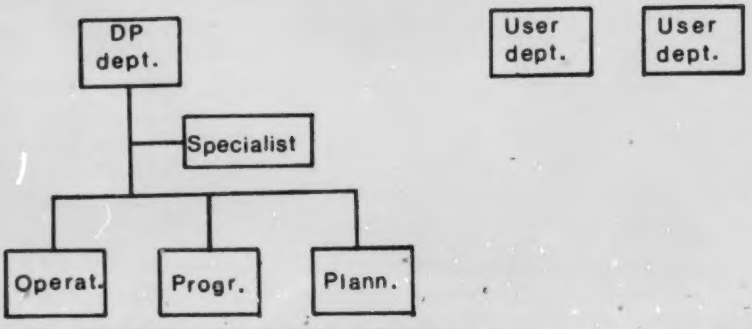

DP dept.
User dept.
User dept.
Specialist
Operat.
Progr.
Plann.

Fig. 12: EDP organization 1965–1970 (functional)

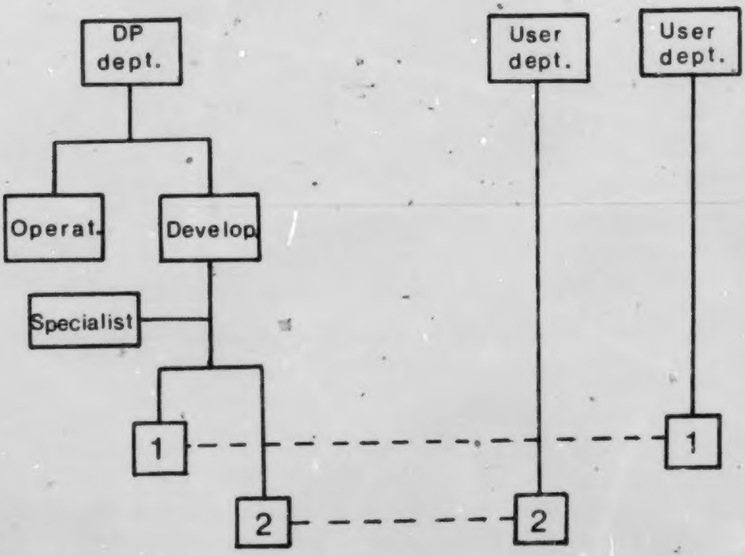

DP dept.
User dept.
User dept.
Operat.
Develop.
Specialist
1
1
2
2

Fig. 13: EDP organization 1970–1975 (matrix)

This situation changed in many places about 1970. Experience showed that new system development was lacking or delayed, and that many new systems did not fulfil the requirements. It was evident that the EDP-organization during its five year old establishment hat somehow lost its original contact with the firm's administrative procedures. In consequence many firms reorganized their EDP-departments, creating project groups with both, planners, programmers, and user representatives (Fig. 13).

The status in 1975 shows that some project groups have worked fine, but many of them have still a rather low performance. An analysis shows two basic reasons: First of all such a project group will normally contain experienced EDP-staff with five-year outdated administrative experience, and experienced user representatives, who have worked with five-year old systems. Secondly this group is confronted with a lot of new und unknown EDP techniques to choose among. It is a general result that too many of these project groups have a tendency to base their work on their outdated experience which leads to frustration. In consequence of this, many firms have reorganized again by creating support functions for the system development work by taking care of economy, technical coordination, and user coordination (Fig. 14).

Nobody knows the future, but management should plan it at least for the period until 1980. I have a personal feeling that the coordination functions and the user part of the system development work will expand, in favour of the traditional system planning and programming functions.

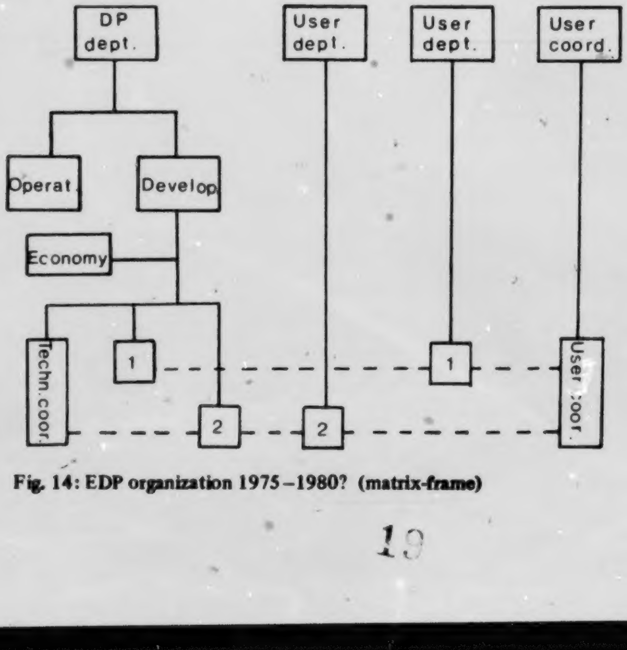


Fig. 14: EDP organization 1975–1980? (matrix-frame)

Standards

The last point I would like to mention concerns standards. The above described organizational development and a formulation of a systems development strategy require manifold standards supporting the EDP-work. It is the nature of management to be objective-oriented, but in the EDP-standards area this has a tendency to create distortion, where the EDP-standards are mainly oriented to the phases of the EDP-project rather than the procedures of the EDP-project. The consequences of such a standards policy is that a great part of the systems development staff knows that they are expected to perform to a much higher degree than the way the work need be performed.

If the above mentioned consequences are to be limited, it will be necessary to create EDP-standards which conceptually are more procedure oriented. Many educational institutions and the more advanced EDP-manufacturers are working within this area, but experience shows that procedural EDP-standards require heavy management support to be effective. On the other hand the management is often reluctant in its support, based on the assumption that procedural standards have not proved results. To solve that problem there is only one way to go – only to implement such procedural standards for which results can be proved in the short range. But the management has to remember that this more reluctant approach to procedural standards places a constraint on the organizational development earlier mentioned. Experience shows that management often forgets this point.

Conclusion

This paper addresses itself both to the staff, who make the systems development work, and to the management who take the decisions about this work. The purpose is to avoid basic management problems by solving them before they become urgent.

The DARC System: A Methodology for Data Bank Organization

Jacques-Emile Dubois, Jean-Claude Bonnet, Jean-François Guaydier, Paris

The DARC system for the handling of chemical data has been described earlier in several papers (see references). Therefore the reader will find here an abstract only.

The DARC system carries out the integration of functions for the handling of chemical data, in particular structural data: description, acquisition, retrieval, computer-aided design (correlation research (1), computer-aided synthesis, computer-aided structural identification).

This account essentially covers the application of the first three functions of the system (description, acquisition, retrieval) for the establishment of a data bank and, in a more general manner, for the problems involved in automatic chemical documentation.

We present the solutions that the DARC system (2) proposes to these problems

1) on the level of concepts and principles concerning the description of chemical structures and, in a more general way, of chromatic graphs. The unique and unambiguous one-to-one description of chromatic graphs is based on the progressive generation of an unique ordered graph by the propagation of an environment which is limited, concentric and ordered (ELCO) (3) and leads to a unilinear topological descriptor which is open: the DEL.

2) on the acquisition level, in other words on the level of the machine input of chemical structures. Automatic acqusition is carried out
- either directly by a specialized microprocessor: the "Topocodeur"
- or by transcoding starting from other codes, in particular from the CAS unique connection tables and the Wiswesser line notation.

3) on the structural data retrieval level. The main problem is the search in substructures. This problem has been solved by setting up the DARC topological screen system. The fragments, reduced to an environment which is limited in B (FREL) (4), make up the heart of the screen system.

As an illustration, the lecture is concluded by the example of two data banks we have implemented, or which are under way towards being implemented: a radioprotective compound data bank (5) and a ^{13}C NMR data bank.

Literature

(1) Dubois, J.E.; D. Laurent; A. Aranda: DARC-System – XVI – Theory of topology-information. J. Chim. Phys. 70 (1973), 1608–1616.

(2) Dubois, J.E.; D. Laurent; H. Viellard: DARC-System description by concentric propagation from a limited environment and modular and nulinear method of representing a symmetrical graph. Paris, C.R. Acad. Sciences, A. 270 (1970), 228.

(3) Dubois, J.E.: DARC system in chemistry. In: W.T. Wipke (Ed.): Computer representation and manipulation of chemical information. New York, J. Wiley (1974).

(4) Dubois, J.E.: French national policy for chemical information and DARC-System as a potential tool of this policy. J. Chem. Doc. 13 (1973), 8.

(5) Dubois, J.E.; D. Laurent; M. Fatome; L. Andrieu; C. Pignalosa: Automatisme. Août (1975).

MAINS Iron & Steel: An Information System for Iron and Steel Materials

Gerd Dathe, Karl-Heinz Dreckmann, Düsseldorf

I. Operation of a Data Base System with Simple Entry Structures

1. Introduction

The Iron and Steel Data Base has been developed by the Betriebsforschungsinstitut at Düsseldorf (Fed. Rep. of Germany). The first considerations date back until 1969 and the first report on the working of the system was given in 1970 (1). More recently, a review has been given in (2).

The system was designed to store and to manage data on iron and steel materials (MAINS stands for: *M*aterial *I*nformation *S*ystem): i.e. data as they are recorded in iron and steel plants. Thus, technical data on products made from iron and steel materials (e.g. rolled metal, wire, tubes) should be collected for various users:
- users of products looking for materials which fulfil certain specifications
- manufacturers of products trying to produce materials according to certain standards
- scientists working in national or international institutions which have the objective to standardize the iron and steel materials

Moreover, the data should be made:
- more transparent
- better and faster available
- easier to manipulate

2. The Problem

Prior to the storage of the data on iron and steel materials in the data base, the data have to be gathered.

This can be done using data sheets which contain the relevant data. The data sheets are produced by the quality divisions of the iron and steel plants from data measured using standardized methods. Those data sheets should describe a particular material completely, if possible.

Unfortunately, this is not possible for the following reasons:
- existing measuring methods will be constantly refined
- new methods will be invented

- old methods will become obsolete.

Moreover, new materials will require new measuring methods. Thus, there is not only one type of a data sheet but several ones, and the single-sheet types will change, as time passes.

Still another way for the gathering of data is to scan scientific papers published in journals and books. To sum up, the data to be stored in the data base can be formatted as well as un-formatted. Even more, the format (i.e. data sheet), if there is one, may differ from one material to another.

So the problem appears to be two-fold:

- the gathering of data has to cope with the multiplicity of data formats (if there are any at all);
- the gathered data have to be stored appropriately in the data base.

3. Outlines of the Solution

To solve both problems at the same time, it seems to be appropriate to use a text documentation system. Such a system allows for the input of textual data in free form and provides for automatic indexing of the text using some sort of thesaurus. The thesaurus would also provide keywords for retrieval purposes. There is only one serious difficulty: the construction of the thesaurus. The thesaurus has to represent the logical relationships between the keywords thus reflecting the context of data items within the document. To give a simple example, the data item "35" on its own does not make any sense. If a second data item is added, e.g. the dimension "N/mm^2", the combination of both data items, i.e. "35 N/mm^2", which has to be interpreted as a data group, provides a little more information. But only the combination with a third element, an identifier, e.g. named "tensile strength", delivers a complete and useful item of information.

Most of the data to be gathered and to be stored in the data base have a character as indicated in the example. But as it is only a simple example, more complicated data structures like (e.g.) periodic groups may occur. So the construction of a thesaurus covering the whole area concerning iron and steel materials seemed not to be feasible at the time the Iron and Steel Data Base was designed (1969).

It was decided to solve the data gathering problem by standardizing the data sheets. A concise set of standardized data sheets was created each of which had a standard structure. Thus, a complete set of data sheets filled in with appropriate data gives a complete description of the actual material. Even the data derived from scientific papers can be filled in the standardized data sheets.

Of course, the set of standardized data sheets is not unchanging: it may be changed by adding new data sheets or deleting old ones. A single data sheet may even be changed to some extent.

To solve the problem of data storage the set of standardized data sheets was simply projected into the structure of the data base. Thus, the structure of the data base may be described as follows:

- Each data sheet is represented by one physical record type of fixed length (a record is meant to be a unit on the physical level).
- If necessary, new record types may be introduced and old ones may be deleted.
- Logically related record types are connected to each other by specific rules which result in logical entries (an entry is meant to be a unit on the logical level).
- Values of records within one entry value can occur multiply per record type; they can also be totally absent.
- Values of entries are of variable length taking advantage of the multiplicity of a record within an entry.

On the level of record types and entry types there exists a description of the data structure (structural data) which is independent and separated from the essential data to be managed (primary data). Thus, application programs become independent of the data stored in the data base.

On the level of record values and entry values there are two mechanisms intended to support the management of the data in the data base:

- Values of records belonging to the same entry values are concatenated. Each entry value contains a set of data items belonging to one material; to each entry an entry number (ENO) is applied which may serve as a substitute for that particular entry.
- Particular data items from an entry or record type may be selected for inversion. The (partial) inversion results in secondary data which is used to provide an associative access to the data stored in the data base (primary data).

While concatenation is useful to support the read function of the data base management system, the inversion is most versatile for retrieval purposes.

4. Working of the System

4.1 Data Gathering

The data sheets which are used for the data gathering, are mainly filled in by the quality control divisions in the iron and steel plants and then sent to the

Betriebsforschungsinstitut. In addition, scientific papers are examined by specialists of the Betriebsforschungsinstitut; relevant data out of these papers are also put onto data sheets.

At the Betriebsforschungsinstitut, all data sheets are checked by specialists for error detection and correction. The check affects not only formal points but also questions of logical nature. After the transformation of the data from the data sheets into punched cards, the cards are checked for errors again.

The transformation of the data itself is performed on a one-by-one basis, i.e. the contents of one data sheet are put into one punched card. The cards are loaded into the data base and, after passing another check, performed by the load program, formed into records (again on a one-by-one basis: one punched card makes up one record). All records belonging to the same material and to the same sample of a product that material are concatenated in order to form a logical unit: an entry Each entry gets (already on the data sheet) a unique entry number (ENO), which is used as the symbolic address of the particular entry for accessing purposes.

4.2 Data Storage

After loading the data base, the records of primary data are stored in two files (Fig. 1.1):
- The Main File (MF) contains the most important records of every entry; the organization is direct using ENO as the key.
- The Overflow File (OF) contains all the other records of an entry not stored in the Main File; the organization is sequential without having a logical sequence.

Those records belonging to the same entry are concatenated using a backward chain. Each chain of OF records has a master record in MF.

Structural data used to describe the structure of primary data, comprise two files (Fig. 1.1):
- The Record Description Table (RDT) is used to describe the structure of the particular records; it contains a statement for each record type of MF and OF.
- The Data Description Table (DDT) is used for the description of the items within a record; it contains a statement for each item type appearing in any record type.

Secondary data are intended to support an efficient management of primary data. Among others, they are formed by inversion lists which allow for an associative access to the data stored in the Main and Overflow Files. There are two

types of inversion lists (Fig. 1.1):
- Index Tables (IT)
- Table of Contents (CT).

The Index Tables are derived from primary data by an inversion of selected items out of the different records (partial inversion). Performing the inversion, the particular item values are not treated as discrete values; instead, they are considered to be members of classes of values into which the whole range of values of an item is divided.

As the Index Tables are subdivided into two parts, there are two files to contain the specific parts of the index (Fig. 1.1):
- IT1 contains the names of those items submitted to the inversion; furthermore, it contains the boundaries of the classes into which the range of the particular item values is divided.
- IT2 contains the actual inversion lists (i.e. lists of ENOs) associated with the particular item names and classes of item values in IT1.

The Table of Contents may be regarded as a special Index Table where inversion is applied to a very important item: the material number. This item is a unique identifier for each material and is therefore part of each entry. The item values are not divided into classes but treated as discrete values.

Analogous to the Index Tables, the Table of Contents is subdivided into two parts; for each part there is a file (Fig. 1.1):
- CT1 contains the discrete values of the item material number.
- CT2 contains the inverson list (i.e. list of ENOs) associated with the particular values of the item material number in CT1

In contrast to primary data, structural data and secondary data are stored in entries of fixed length (i.e. one entry consists only of one record). All files are sequentially organized; the entries are sorted in an ascending order. Logically related entries within the files DDT, IT2, and CT2 are concatenated by a single chain and each has a master in the files RDT, IT1, and CT1, respectively.

Besides structural data, primary and secondary data, the data base contains two more types of data: auxiliary data and working data. These types of data will not be discussed; only the Entry File (EF) will be mentioned (Fig. 1.1) which is used for the management of the storage space in the Main and Overflow Files.

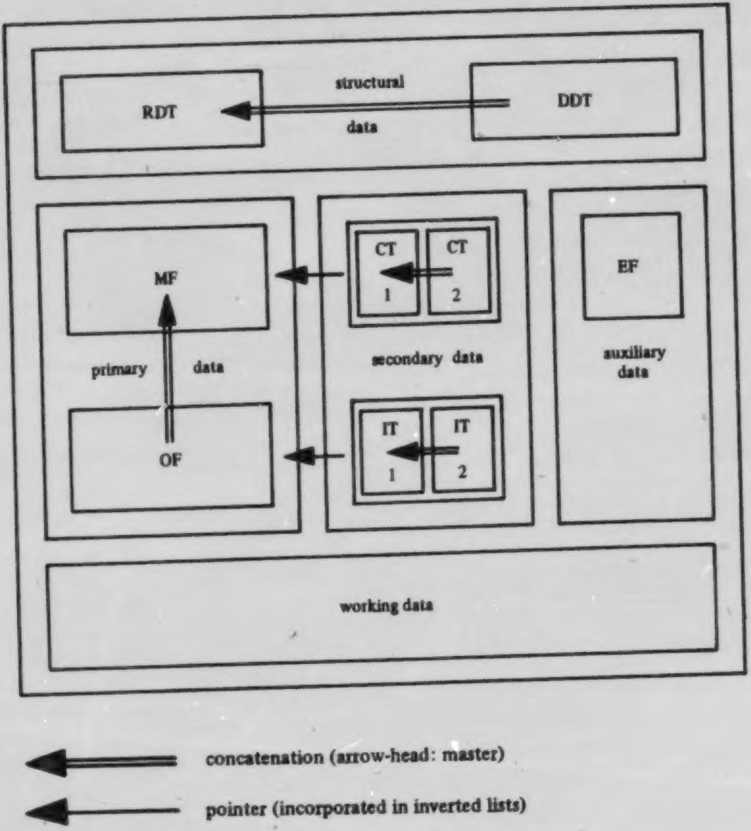

structural
data
RDT
DDT
MF
primary
data
OF
CT
1
CT
2
secondary data
IT
1
IT
2
EF
auxiliary
data
working data
concatenation (arrow-head: master)
pointer (incorporated in inverted lists)

Fig. 1.1: Structure of the iron and steel data base

4.3 Queries

The data base management system which serves to manage the data stored in the Iron and Steel Data Base, has a modular structure. It comprises a number of programs and subroutines which can be subdivided into three groups:
- programs for the preparation of the data base
- programs for the update of data within the data base
- programs for the retrieval of data from the data base.

To prepare a data base, the following tasks have to be performed:
- Storage space for the data base (i.e. Main and Overflow Files) has to be allocated on peripheral memory; the Main and Overflow Files have to be initialized.
- The Record Description Table has to be established.
- The Data Description Table has to be established.

Preparation of a data base is not only performed for the initial creation of a data base, but also for its extension in the course of the running of the data base system.

For an update of data within the data base, one or more of the following tasks have to be done:
- storage of data in the data base (Main and Overflow Files)
- deletion of data in the data base (Main and Overflow Files)
- setting up, extension or deletion of Index Tables
- setting up or deletion of the Table of Contents
- updating of Index Tables or of the Table of Contents

The retrieval of data from the data base comprises a search and a read process. Both processes are of great interest for every user of the data base system. Retrieval can be done in two ways:
- The symbolic addresses (ENOs) of the wanted entries are known. In this case, the specific entries can be read randomly from peripheral memory.
- The symbolic addresses of the entries are unknown. This is the more interesting case because the addresses have to be found during the search process.

The search process is performed via an associative mechanism. It is initiated by formulating a search profile which serves as input to the retrieval program (Fig. 1.2). Such a profile consists of three parts:
- First of all, a list of the names of those properties (items) is given, the values of which are wanted by the user (wanted properties).
- Then a restriction is set up containing the names of those properties (items) which are already known. The item names are connected via logical operators (AND, OR, AND NOT, OR NOT).

- Finally, the values of the items within the restriction are given by minimum and maximum values (given properties).

On the basis of the restriction and the given properties, the symbolic addresses (ENOs) of those entries are found which fulfil the search profile. This is done in the usual way: for each given property the appropriate Index Table or the Table of Contents is scanned extracting from the table the suitable entry numbers. An entry number is suitable, if the associated entry contains a value of the specified item which lies within the Min-Max range of the corresponding given property. Thus, for each given property there results a set of suitable entry numbers.

Test temperature			wanted properties
Tensile strength			
Heat treatment			
Material & C & Si			restriction
Material	Min (1,5415)	Max (1,5415	given properties
C	Min (0,12)	Max (0,20)	
Si	Min (0,15)	Max (0,35)	

Fig. 1.2: Example of a search profile

A match process according to the restriction is applied on these sets of suitable entry numbers in order to achieve the final set of resulting entry numbers. As the symbolic addresses of the entries which fulfil the search profile are known now, the specific entries can be read randomly from peripheral memory.

The read process is performed on the basis of the entry number (ENO) of the entry to be read. From that number the physical address of the associated entry in the Main File is derived and then an access to peripheral memory is performed randomly using this address. Thus, having read all records from the Main File which belong to the entry, the associated chain of records from the Overflow File can be read serially following the chain.

5. The Use of the Data Base System

The results of the search and read processes performed by the data base management system are stored again on peripheral memory. They can be read from there by any application program and then be processed in an appropriate manner. Standard application programs are:

- Editing.
  Controlled by parameters a number of different reports from the data base can be printed.

- Statistical evaluation.
  The data within the data base may be evaluated using a set of statistical methods.
- Plotting.
  Results of evaluations can be transformed into control statements for a plotter and then be displayed graphically.

II. Design of a Data Base System for More Complex Data Structure

1. Introduction

Many efforts have been made to develop generalized Data Base Management Systems (DBMS). Progress can be recognized even though the final goal is not yet defined in a widely accepted manner. This is also true for a generalized data model which is a prerequisite for a generalized DBMS.

Therefore it is not surprising that problems can arise to which existing DBMS are not applicable. Such a problem was encountered when dealing with data derived from standards in the field of iron and steel materials. These data are recorded in tables on printed paper and are rather difficult to handle automatically for the following reasons (among others):

- Each table as a rule is of a type of its own. There are at least 500 tables in the beginning (German standards) and probably many more when the data are augmented by foreign and international standards. A transformation from these many different table types to a reasonably lower number seems to be unfeasable.
- The retrieval of a standard material based on a given set of values of properties has to be completely non-procedural because there is known neither which data type contains standardized properties according to the given set of properties, nor whether a given property is standardized at all.

For the solution of these and some other problems a new DBMS has to be developed. This system has features that may be of general interest if the problem is tackled of extending system capabilities or implementing data models not inplemented until now. Other features result from efficiency considerations for the special application and may be modified for other boundary conditions.

Having the goal of developing a DBMS within a reasonable time it seems appropriate to rely on a data organisation which is simple, commonly accepted and which can form the basis for more complex data organizations. This organiza-

31

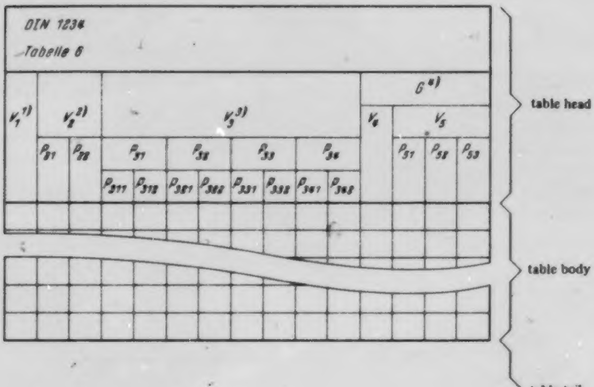
tion might be that of entries (here used as a synonym for logical records) and blocks (here used as a synonym for physical records).

One of the major problems in the field of DBMS is the management of data more complex than an entry. The considerations in chapter 3 concerning this topic may be of interest for those discussing data models.

2. The Problem

From a DBMS end-user's point of view the data structure is given by the structure of tables as shown in Fig. 2.1.

When transforming this structure into an entry or an association of entries the contents of a matrix element in the table body is a candidate for an item value. The meaning of the contents of a matrix element is indicated by the heading of the corresponding column in the table head. The heading of a column can be interpreted as an item name, though it contains more information than an item name usually does: It can be composed of names for variables V (e.g. tensile strength in N/mm2) and of parameters P (e.g. temperature = 20°C). Note that a parameter is a combination of a name and a value. All the tables in the German standards for iron and steel materials contain about 50 variables. Their



1) variable (represented by its name)
2) variable with one parameter (which has two values)
3) variable with two parameters
4) group consisting of two variables.

Fig. 2.1: Structure of a table

combination with an arbitrary number of parameters results in a practically unlimited number of different column headings especially when the system is extended to foreign standards. Column headings can build up a group G (e.g. mechanical properties).

By another part of the table head the table as a whole is identified. The table tail contains footnotes which are quoted in the table head or in the table body. A row of the table body is a natural logical unit which therefore is considered as an entry. This choice has some advantages over considering a column or a whole table as an entry: A row will frequently be the object of output; a row is not too complex in structure; all rows of the table body of an individual table are equal in length. Other reasons are due to the efficiency of the retrieval process.

There is no unique identifier for a row in the tables. It is true that each row contains a standardized material number but sometimes a row contains several numbers or several rows of an individual table contain identical ones.

The data out of the table head are needed for interpreting the values of a row as well by the user as by the DBMS. These data will be stored in entries of a type different from that of the row entries. A third format is used for footnotes. The complete instance of all table entries is derived from the information in the table head, the table body, and the table tail. It is called an association.

There is another association of great importance: the standardized properties of an individual material are scattered over an unknown number of tables. The association of all rows for a certain material from those different tables often is the object to the find process rather than isolated rows.

The generation of data output can be divided into two distinct stages: find and get. The find function locates entries and associations of entries for certain materials, the properties of which fit into a given find profile (Boolean expression containing wanted properties). Sometimes other associations representing instances of certain tables or standards may be wanted to be found, too. The find profile is a non-procedural input into the find function. The DBMS has to transform the find profile into an adequate procedure, an essential part of which is the selection of entry types addressed by the find profile. The results of the find function are identifiers of entries or associations which may or may not be handed over to the user. The identifiers are the input into the get function.

Data from located entries or associations are read by the retrieve function. Names of items to be retrieved are specified by a retrieve profile. Item names correspond to column headings which are also part of a property in the find profile. However, since the find and retrieve functions are separated from each other, the item names incorporated in both profiles may be quite different. It is not sure that specified item names are provided for the specified entries at all. Therefore in many cases it is necessary that means for interaction between the user and the system are provided.

Approximate quantities of German standards are 500 tables, 2000 materials, 1.000.000 characters. The data on a certain material are scattered, on average, over 5 tables.

3. Some General Aspects of Entries and Associations

An entry may be defined as a data aggregate which a program can read or write in one single step once an identifier of the entry is known. Such an identifier may be given explicitly (e.g. by an entry key) or implicitly (e.g. by a statement like "next in sequence"). So it should be clearly recognized that an entry is not defined on its own but only in connection with a program.

As there exist various types of programs an entry may be more or less specific for the particular program to which it is connected. On the one end of the scale there are application programs designed to serve single users. Therefore the entries connected with such an application program usually are tailored to the needs of that particular program and the associated problem. On the other end there are channel programs within the operating system designed to serve a great variety of programs for all kinds of problems. The entries connected with system programs are therefore tailored for general use within one EDP machine and one operating system.

An application program now is just one level in a hierarchy of programs with other programs on lower levels (macros, subroutines, utilities) for the communication with the peripheral memory. The channel programs are situated on the lowest level and perform the actual reading and writing on the peripheral memory. For this purpose channel programs often use standardized entries of fixed length which are called blocks or physical records. Entries used in application programs in general are not standardized at all. So for the storage of data on peripheral memory there has to be a transformation of non standardized entries to the standardized ones which can be handled by the channel programs. This transformation is done when appropriate macros, subroutines or utilities of the operating system are called by some application program to finally initialize certain channel programs.

In the case a DBMS is used it represents a new level between the application program and those parts of the operating system which accomplish I/O-operations like macros, subroutines, utilities and channel programs. The entries connected with the DBMS play a very important role in the overall system consisting of application program(s), DBMS, and operating system: the entries used in the DBMS separate programs as represented by the application program(s) and data as stored on peripheral memory. To do this in a most adequate and efficient way there has to be a data description language to describe the structure of the DBMS entries.

If the structure of the DBMS entries is changed to serve new or modified application programs there is no need to change existing application programs as long as the new structure of the DBMS entries is compatible with the old one, i.e. application programs are no longer coupled via stored data they use in common. Moreover, the assignment of DBMS entries to blocks can be changed according to efficiency considerations without affecting any application program.

So the use of a DBMS separating program and data results in two transformations instead of one, the first one concerning the entries of the application program(s) and the DBMS (as above), the second one concerning the entries of the DBMS and the operating system (i.e. channel programs). To perform these two transformations in a way nearly optimal the data description language hat to be very flexible and besides that, there has to be a quite powerful data manipulation language.

To achieve some flexibility a DBMS should be able to compose an application program entry from parts of one or several DBMS entries. Application programs themselves usually divide their entries into parts which are objects of arithmetic operations etc. Two more data aggregate types are sufficient for this purpose: items and groups. An item is the elementary data aggregate which is not further divisable. A group is built up from items and groups which are treated as a whole for some purposes. So by the recursive definition of the group an entry may be regarded as a group with the special feature that it can be interchanged between programs via read and write functions.

An association may be defined as a data aggregate which consists of entries and the interrelationships which are known to a program. Examples of associations are:
- the (logical) file of ADABAS (3)
- the data base of IMS (4)
- the group relation of the CODASYL Feature Analysis (5)

- the owner-coupled set of the CODASYL-DBTG (6)
- the set in a mathematical sense

There are many more types of associations, e.g. so called segmented records, areas, and inverted file or pointer coupled entries. In fact, any collection of entries may be interpreted as an association if the interrelationships between the entries are known to a program. The nature of this knowledge may differ but it has to be present in one form or another. So again, as in the case of entries, associations are not defined on their own but only in connection with a program.

The representation of an association depends on the representation of the interrelationship between the entries of the association. This interrelationship may be represented by

- position of the entries (entries in physical adjacency)
- pointers (concatenation, inversion, etc.)
- contents of the entries

Assuming again three levels of programs (application programs, DBMS, and operating system) the representation of the interrelationship by contents of the entries is typical for application programs. In general, many entries have to be inspected to find those which belong to a certain association – many more than really belong to it. This can be avoided if provisions have been made to transform the interrelationships represented by contents to interrelationships represented by pointers or by position. In a more complex application an individual entry may participate in several associations, but only for one of those associations can the interrelationships be represented by position.

The DBMS is a link between the application programs and the operating system to provide a means for an effective and uniform representation of associations in application programs. This is done by appropriate features of the data definition and data manipulation languages.

On the transition from one level to another an entry on one level can be part of an association on the other level and vice versa, if associations are involved at all. An application program entry e.g. could be part of a DBMS association and just the interrelationships of this association could be represented by position within a single channel program entry (block).

4. Overall Design Philosophy

A DBMS may be considered as a part of a data base supported information system (fig. 2.2). Such a system consists of hardware and software and provisions are made for man to communicate with the system. Looking at the software only four major parts can be recognized:

- The data base management system is a standard software system designed to manage data centrally and uniformly.
- The data base comprises the data which is managed by the DBMS.
- An application program is a program which interacts with the DBMS in order to solve a particular problem by calling functions of the DBMS. The application program and the DBMS communicate via a generally accessible interface. DBMS used in this way are often called host language systems.
- A standard application program is an application program which is commercially available and support more general applications like report generation and query language interpretation. For efficiency reasons those programs often use specialized, not generally accessible interfaces. An undivisible combination of a DBMS and a standard application program may form a self-contained system.

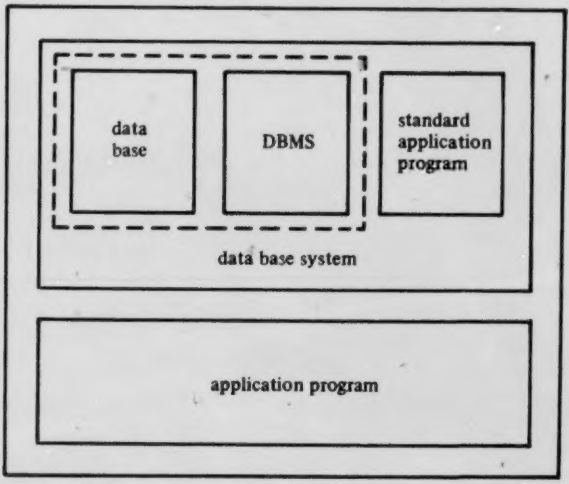


Fig. 2.2: Data base supported information system

The data base management system and the data base can be looked at as an unit which then may be called the (basic) data base system. (Sometimes the basic data base system and a particular standard application program build up an undivisible unit which is also called data base system). Such a basic data base system without any application or standard application program is the design goal and will further on be refered to as the data base system.

The data base system will have a clearly defined, generally accessible interface which is used by an application program via subroutine calls rather than via an extension of a programming language. This interface has to be used by a dialogue program which is the main application program to be developed in the future. So the whole information system tends to be a question answering system.

5. Features of the Data Base System

5.1 The Data Base

One of the most important steps in the process of developing a data base is the design of the data structure. Three aspects have to be taken into consideration:
- Structure, quantity, and processing of the original data,
- capabilities of the DBMS for structuring the data of the data base, and
- costs for storing and processing data structured in different ways.

In a data base generally five types of data can be distinguished:
- Primary data:
  primary data represent the original data.
- Secondary data:
  secondary data are derived from primary data in order to support the functions of the data base management system.
- Structural data:
  structural data describe the structure of the data in the data base as far as it is not intrinsic.
- Auxiliary data:
  auxiliary data are used for internal administration purposes.
- Working data:
  working data represent intermediate results which are local to the DBMS.

For the transformation of the original data into the primary data some data aggregate within the original data has to be regarded as an entry. The appropriate data aggregate is a row in a table, as mentioned above. This decision results in a great number of entry types in the data base as there is a great number of tables in the original data and each table requires at least one entry type of its own. So this high structural diversity is one of the major problems in the design of the data base system.

In order to solve this problem structural data are introduced. Structural data may be considered as the stored result of a translation of statements of a very powerful data description language. This data description language itself is used to describe each table in its particular structural properties. The structure of the structural data itself may be intrinsic or may be put in with the help of a syntax-directed generator, thus considerably increasing the flexibility of data structure descriptions. This method seems to be capable of describing the structures of all kinds of tables and of similar data aggregates, not to speak of simpler data structures.

Structural data are utilized in the find and retrieve processes as they are needed to interprete the stored primary data. As far as sequential find and retrieve processes are concerned, no further support for those processes is needed. But since sequential find and retrieve processes do not suffice at all for a data base systems there has to be support for random access to the primary data.

One of the most versatile techniques to support random access to primary data is the inverted file technique. It must be distinguished between a complete inversion and a partial inversion of the primary data. In the case of the given data material it seems appropriate to invert those data partially only for the following reason: For variables without parameters the inversion mechanism functions as usual and straight forward. But as soon as parameters are present the inversion mechanism becomes rather complicated and expensive. So the inversion will be applied only to variables without parameters.

This inversion mechanism is called Inversion I. If this mechanism is applied to a variable, the values of the variable are listed together with the symbolic addresses (entry numbers) of those entries in which the particular values occur. But besides this inversion mechanism a second and somewhat new inversion mechanism can be applied to the variables of the primary data, which lists the names of variables together with the symbolic addresses (table numbers) of those tables in which the particular names occur. This kind of an inversion is called Inversion II and it may be applied to all variables, no matter whether parameters are present or not. It might even be applied to the parameters themselves.

The variable names are contained in entries of the structural data which are used to describe particular tables. The collection of all structural data entries describing one table gets the number of that particular table. Therefore Inversion II gets an interesting interpretation: It may be interpreted as an inversion of the structural data while Inversion I is the inversion of the primary data.

Inversion II is useful not only in cases when Inversion I is not applicable but also in all those cases when the find and retrieve processes are not directed towards an entry or a set of entries but towards an association or a set of associations which may happen quite often. The following DBMS associations of major importance can be distinguished:

- All the primary data entries belonging to the same table constitute an association. For flexibility reasons the interrelationship between the entries of this association is representend by pointers. In addition, to increase the efficiency in handling complete tables the entries are usually placed in physical adjacency, but the DBMS does not rely on this property.
- Entries containing identical variable values form an association for a particular value, each. A very important example is the association based on identical material numbers. The interrelationship of this special association is represented by pointers (Inversion I). As mentioned above, Inversion I is not feasable for all variables.
- Entries containing data which belong to identical variable names form an association for a particular variable, each. The interrelationship of this association is represented by pointers (Inversion II).
- All entries fitting into a find profile form an association. The interrelationship of this association is dynamically defined by the find profile and, in general, represented by contents.

As auxiliary and working data are of minor importance compared with primary, secondary, and structural data they will not be discussed in detail. They are mentioned for completeness reasons only.

All types of data in the data base are treated in the same way which may be outlined by the following principles:

- Each DBMS entry gets an internal entry number which is automatically assigned by the DBMS.
- The internal entry number serves as a symbolic address of the entry.
- All DBMS entries are of variable length.
- A variable number of DBMS entries is stored in a block (channel program entry) of fixed length.
- The storage of data on peripheral memory is independent of particular memory units.

5.2 The Data Base Management System

Another major task in the design of a data base system is the development and implementation of functions of the DBMS which may be divided into three groups:

- Basic functions:

  Basic functions of the DBMS are the functions for the manipulation of data (find, get, modify, delete, and insert) and for the support of the integrity of the data (data security and data privacy).

- Additional functions:
  Additional functions of the DBMS are such functions which are needed for an efficient operation of the data base system (e.g. reorganization of the data base).
- Auxiliary functions:
  Auxiliary functions of the DBMS are used in connection with basic or additional functions to solve smaller tasks (e.g. conversion of data).

As the find and retrieve functions are of outstanding interest the following will concentrate on these two functions.

The find function has to locate entries or associations of entries in the primary data according to a given find profile, as mentioned above. A find profile is a Boolean expression containing all information which is necessary to achieve the location of the wanted entries or associations. Every find profile can be divided into parts, each one of which is associated to a certain variable by giving at least a variable name and, in many cases, a relational operator and a variable value. All those parts of a find profile are linked together by boolean operators.

The data base management system now transforms the non-procedural find profile into a procedure which has the following major steps:

- Step 1:
  According to the variable names in the find profile those entry types in the primary data are located which contain values for these variables. This is done via inverted lists derived from Inversion II. It should be noted that an association in the primary data representing a table is connected to an association of structural data entries which describes the table as a whole (see above).
- Step 2:
  Depending on whether the single variables have been submitted to Inversion I, one of the following two steps have to be carried out for each variable.
  Step 2.1:
  If a particular variable has been submitted to Inversion I it is easy to obtain all the entry numbers of those entries which satisfy the appropriate part of the find profile from the associated inverted list.
  Step 2.2:
  If a particular variable has not been submitted to Inversion I the related associations in the primary and the structural data for those tables which cover the variable have to be put into main memory. The particular entries

41

of each primary data association are sequentially scanned under control of the related structural data association in order to catch all entries which satisfy the appropriate part of the find profile.

- Step 3:
  The result of step 2, in general, is a number of lists of entry numbers, just one list for each part of the find profile. These single lists have to be condensed to one list according to the Boolean operators between the parts of the find profile. This is a more complex process in cases where several tables are concerned. In such cases several transformation and normalization processes have to be carried out before achieving the final result.

The set of entry numbers which is the result of the find process may be interpreted as the pointer representation of the interrelationship of an association, which has been dynamically created during the find process. This is contrasted by a static association for which the interrelationship between its entries is represented physically in one way or another (see above).

This dynamic association now is the input to the retrieve function which actually accesses the entries which belong to the association. Additional input is given by a retrieve profile which specifies the items and groups which shall be read from the entries located by the find function. Besides that the retrieve profile specifies the output format of the specified items and groups. It should be noted that the retrieve function is not a trivial task, for the following reasons:

- find profile and retrieve profile may be different according to the needs or wishes of a particular user,
- in general, more than one table is involved in the retrieve function, and
- the user may want to get more or fewer entries than really belong to the dynamic association.

The last wish can only be fulfilled in the course of a dialogue between the user and the system.

In some cases, which could occur quite frequently, the dynamic association which serves as input to the retrieve function might be a static association or a set of static associations. So this points it out clearly that it would be very helpful to provide the retrieve function with options which are directed towards associations.

When data are read into the main memory by the retrieve function they may be processed in an arbitrary manner or handed over to other basic functions of the DBMS, e.g.

- Modify:
  The modify function changes an entry in the primary data.
- Delete:
  The delete function erases an entry from the primary data.
- Insert:
  The insert function integrates a new entry into the primary data.

In all cases, the secondary and auxiliary data have to be updated simultaneously.

6. The Usage of the Data Base System

The data base system the main features of which are outlined above is designed to be used in connection with different application programs. The most important of these programs will be a dialogue program which has to be developed in the future. The overall system then comprising the data base system and the dialogue program shall make up a question-answering system.

This system will work in a multi-user mode to allow for more than one dialogue with the system. Moreover, it will be able to perform process batch and dialogue tasks simultaneously. Finally, the whole system is designed and implemented in a way to make it portable from one EDP machine to another.

III. Outlook

The two systems described in parts I and II are independent of each other; they can be operated as two separate question answering systems in their own rights, when the new system will be implemented and ready to start working.

Meanwhile, the current status of the systems is as follows:

- The Iron and Steel Data Base has been running very well for a couple of years. The volume of the stored data is growing steadily: the data base contains more than 11 300 entries on 325 different materials (dead-line: Jan. 1, 1975). A set of application programs for editing, statistical evaluation, and plotting of data out of the data base has been developed. A special application program will be developed in the future, which allows for interaction with the data base via a CRT terminal (dialogue).
- The new system has been designed in a way allowing for a usage of the fully implemented system in many application areas; therefore it may be called a generalized data base management system. The design has been effected following the principles of Structured Programming (or rather Structured Design) resulting in a voluminous design documentation. Implementation of the system has begun July 1, 1975; the implementation will again be

carried out according to the principles of Structured Programming.

A very interesting point should be noted: in order to make the system portable, the macro processor STAGE2 will be used. With the help of STAGE2 an implementation language will be developed tailored to the needs of systems programming, especially to that of implementing a data base management system.

A first version of the new system will be operational by the end of 1976; it will comprise most of the nucleus of the system and a few utilities. The rest of the nucleus and most of the utilities will be realized in the following years.

When the new system is ready to run, it will run in parallel to the old system (the Iron and Steel Data Base) for quite some time. Programs will be developed performing a bridging operation between the two systems. Later on, the new system will probably take over the functions of the old system thus making the old one obsolete.

The new system is described in more detail in (7) and especially in (8). Finally, a new software system should be mentioned, which has been developed in connection with the new data base system: DACAPO – a system for *d*ata *a*cquisition from a *c*omplexly *a*rranged *p*ool of data (9). This system was designed in order to provide the data to be stored in the data base.

The use of DACAPO however is not limited to data acquisition from standards for iron and steel materials; the system is applicable to all data which have complex structures (not to speak of data with simple structures). Moreover, the system which is run in a dialogue mode via a CRT terminal, is portable and can therefore be installed on different computers.

The development of DACAPO is completely finished; the system has been working for some time.

Acknowledgement

The development of the data base systems has been supported by the Institut für Dokumentationswesen, Frankfurt, Fed. Rep. of Germany, since 1971 using grants by the German Federal Ministry of Research and Technology.

Literature

(1) Mommertz, K.H., J. Radestock, K.J. Kremer: Entwicklung einer Datenbank als Informationssystem über die Eigenschaften von Stählen. Stahl u. Eisen, 90 (1970), p. 1339–1346.

(2) Born, A.: The Iron-Steel Data Base as an Example of a Data Base Scheme. Paper presented on the symposium on: Man-Machine Communication for Scientific Data Handling. Organized by the Task Group on Computer Use of the ICSU Committee on Data for Science and Technology (CODATA). Freiburg im Breisgau, July 22–27, 1973.

(3) Schnell, P.M.: ADABAS – Adaptable Data Base System-Introduction. Software AG, Darmstadt (1972).

(4) Information Management System/360 for the IBM System /360 (System Description) Version 2 – General Information Manual.

(5) CODASYL Systems Committee: Feature Analysis of Generalized Data Base Management Systems (1971).

(6) CODASYL Data Base Task Group: April 71 Report (1971).

(7) Dathe, G.; K.H. Dreckmann: Entwurf eines Datenbanksystems für normierte Kennwerte von Eisen- und Stahlwerkstoffen. Gesellschaft für Informatik – 4th annual meeting, Berlin, October 9–12, 1974. Lecture Notes in Computer Science, 26 (1975), p. 484–493. Berlin-Heidelberg-New York: Springer.

(8) Dathe, G.; K.H. Dreckmann: G. Henselmann: Design Principles of a Data Base System for Unlimited Quantities of Data with Highly Complex Structures. Paper submitted to: International Conference on Very Large Data Bases, Boston, Mass., September 22–24, 1975.

(9) Dreckmann, K.H.; G. Hofmann: Ein Programmsystem zur Erfassung von Daten aus komplex strukturierten Tabellen. Paper presented at: Gesellschaft für Informatik – 5th annual meeting, Dortmund, October 8–10, 1975.

Statistical Methods in Data-Documentation1)

Wilhelm Gaus, Ulm

There are several contrasts between literature documentation and data documentation. Let us pick up two of them to show why statistical methods are much more important in data documentation than in literature documentation.

First, a document or more generally a unit of documentation is in literature documentation a book, an article or something like this, in data documentation it is mostly a single value or property. It is much easier to collect (let us say) a million data (this is e.g. the annual output of a medium scaled clinical laboratory or the input in one second to a computer monitoring a fast process) than to collect a million publications. So generally a data documentation system contains many more items than a literature documentation system.

Secondly data can easily be aggregated, so we distinguish between single data, data sets, and mass-data. Often the information finally required is represented in a mass of data. Let us take an example: The wind speed and direction at a certain airport at a certain time is important as a single data for the pilot of an aircraft landing at this airport at this time. The same measurement is as an isolated value nearly without any interest for a weather forecast office. There a set of wind speeds and directions from several locations and maybe several times is required. So the same data item can be both, a single isolated value and an element of a set of data, depending from the user's interest.

Furthermore, the same data elements can join different sets, depending again on the user and the purpose of his retrieval. As an example from medicine let us take the blood pressure of a certain patient at a certain time under certain conditions. This value joins the set of data of this patient, containing besides the blood pressure, the weight of this patient, his temperature, the results of several lab-tests, etc., and it is used for diagnostic and therapeutic purpose for this patient. The same blood pressure joins another set of data too, which is formed for medical research, e.g. all blood pressures from a collective of patients,

---

1) Prof. W. van Eimeren from the Department for Medical Information Processing, Statistics, and Biomathematics at the University of Munich, who presented a lecture with this title at the workshop, found no time to prepare a written version of his presentation. Because I know how important are statistical methods in data-documentation and because this field is not very well known, Prof. v. Eimeren and I agreed that I incorporate a contribution with this title in this volume. W.G.

maybe detailed in subsets for age-groups, diagnosis, measurement conditions etc.

The usual statistical methods and techniques can easily be studied in statistical textbooks and journals. Therefore we will not deal here with the standard statistical procedures, but stress their application and usage in data documentation and the helps they can give us there. Statistical methods are employed for a lot of tasks within data documentation. We will limit ourselves in this paper to five major and widely used topics.

1) Data control

Each data item entering a data documentation system should be checked for correctness, plausibility, and accuracy as far as possible. Let us consider a data bank holding one or more sets of data. Then the information contained in a set of data can be used to check whether a certain new-coming data element fits into the set. If it doesn't the following reasons are possible:

- The original data was wrong encoded, wrong transferred, or wrong decoded or the data item was by mistake announced to be a member of the set but does not belong in reality to this set. For the latter case the following examples are given: A value is labeled as blood pressure but is the pulse frequency, or a value is labeled to belong to Mr. X but is in reality a value of Mr. Y.
- The data generating process is out of order and delivers therefore wrong data, e.g. the laboratory work is not done well, a part of the production machinery works incorrect, or a wrong sample or material was tested. With this important situation the data documentation challenges the generating process as a quality control system.
- The rejected data element is a true outlier, it belongs to the set, and the alarm of the control was a false one. This can happen if the set of data already collected has only few or homogenous elements, if the control mechanism is too strict, or if the new-coming data brings a lot of new (and therefore important) information.
  Additionally, if the data generation is considered as a stochastic process it may happen – of course with an extremely small probability – that true extreme values occur. In this case only a higher frequency of extreme values justifies rejecting them.

Values which are rejected by a statistical data control procedure are in fact mostly wrong data. At least the probability that data are rejected although they are correct (the so called α-error) can be computed exactly in most situations. But we should keep in mind that a value not rejected by a statistical control procedure may not be a correct one. It is a matter of fact that the power of the statistical procedures to detect wrong data (the so called β-error = probability that a wrong value is not rejected) is very difficult to determine and usually we know only a rough estimate.

Besides the simple checks, whether a value is within a given range or if there is an illegal combination of properties, control charts are the most often used statistical technique for data control. In control charts (for one or more variates) the probability is determined that a value like the new-coming one occurs in a given distribution (most normal-distribution). The value is rejected if this probability is beyond a certain limit.

2) Data presentation

Often the result of a retrieval in data documentation is not a single datum but a set of data. Of course the user should look at each element of the retrieved set. But frequently the most striking information is given in a total view of the set. Therefore all the methods and techniques of descriptive statistics such as tables, contingency tables, graphs, uni- and bivariate plottings, index numbers, etc. may be employed to present the information which is contained in the retrieved data. Moreover, the importance of data presentation with the techniques of descriptive statistics is stressed because we know that there are a lot of possibilities "to lie with statistics". "To lie with statistics" means not to offer objectively wrong data or results, but to present the retrieved data in tendentious way.

3) Data compression and data reduction

Sometimes it is not worth keeping all data for a longer time from a process generating a big mass of data. An extreme example is the data set which springs from the conversion of an analog signal into digital data. In a lot of applications it is better to enter only pulse-height and pulse-frequency into a data documentation system. Besides the analog-digital conversion there are many other applications for data compression and data reduction. In many cases it is fully sufficient to store the frequency distribution and its characteristics such as mean and variance instead of all the data. Sometimes it is enough to store the result of a statistical analysis and the original data can be dropped.

It must be stated however, that data reduction is always a loss of information on one hand and a saving of storage capacity and retrieval time on the other hand. So we have to weigh the arguments and make a decision for each data base. If we decide for a data reduction at input this decision is irreversible at retrieval time.

Data compression and data reduction can not only be used at the input into a data base, it can be done during the retrieval too. Whether data compression during retrieval – or more precisely as a second task after the retrieval – is done or not depends on the user's needs and interests. In literature documentation the searcher will ask the user if he wants to have a better recall or a better precision. Correspondingly the searcher in a data documentation system has to ask his user if he wants the original data or if he wants the data compressed and reduced and, if either, how he wants it reduced, e.g. frequencies of properties, range, mean and variance, rank order, the appropriate statistic for hypothesis testing to a given question or the comparison of subjects, a specific estimator, etc. Obviously there are some points without sharp boundaries between data compression at retrieval time and the techniques of data presentation.

4) Generalization and prediction

A lot of data banks contain data from surveys, investigations, treatments, etc. not taken from a total population of persons, animals, plants, business firms, etc., but from a sample out of such a population. The survey, investigation, treatment or whatever was originally done to come to results and statements not only on this sample but for the underlying population. Therefore the retrieval has to be enlarged with the statistical methods of estimation, extrapolation, prediction, and prognosis. Somebody could have the objection that these are separate tasks and not parts of a data documentation system and formally he would be right. But both the data documentation and the statistical generalization and prediction use the same tool, namely the computer, and the same raw material, namely the original data. Therefore the statistical procedures for generalization and prediction, which are formally not an original part of data documentation, are mostly integrated into the retrieval procedure.

One situation should be mentioned especially. Several data banks contain information (e.g. business data) which is gained again and again after a certain time, e.g. every year. The data are routinely scanned for trends and used for prediction of the development in future, e.g. the coming year. Here data acqusition, data documentation, and statistics are highly integrated into one system.

5) Search for data imminent structures

The main purpose of some data banks especially in science and research institutions is not to retrieve single data items or data sets, but to look for unknown structures and patterns in the stored data. Such structures are correlations, dependencies, clusters, common or specific factors of influence, trends, or mathematical relationships of various types and are mostly obscured by random effects. To detect such structures and to describe them, the techniques of information retrieval, ways of data manipulation, and statistical methods have to be put together into one strategy.

Let us hope that the above briefly mentioned topics are enough
- to demonstrate the necessity to use statistical methods for a powerful data documentation
- to give some ideas how statistical methods are involved in data documentation, what their tasks are, and to what type of results they lead
- to show, that a statistician is – among documentalists, EDP-men, and the professionals from the subject documented – an essential member of the crew developing and driving a data documentation system.

Clinical Laboratory Data, Capture and Evaluation

Klaus-Peter Killian, München

1. Introduction

The demand for laboratory work especially in the fields of preventive medicine and intensive care has increased the daily number of examinations in biochemical laboratories. Furthermore the advances made in the field of quantitative examination methods have given the doctor many more possibilities in the diagnosis of diseases. Therefore the number of tests in clinical laboratories increased exponentially within the last five years. This trend cannot be mastered by the given organization of the existing laboratories. These tasks can only be carried out economically and efficiently by use of an electronic data processing system. Such a system has been developed under the direction of the West German Federal Ministery of Research and Technology in the hospital of München-Harlaching. With this system, operating since 1970, the costs for one test could be reduced from DM 2.13 to DM 1.44.

Three fundamental problems must be solved when implementing a computer system in the process of work and organization of the existing laboratory;

1. Mechanization of the laboratory by using instruments based on analytical procedures, which require only small quantities of material for examination, and only a few manual steps.
2. Simplification of the work procedure through the convenient division of the individual stages, by avoiding delays and by cutting out complete manual work steps, such as fractioning of samples. The Silab system, created in collaboration with Siemens, allows one to reduce the amount of calculation by processing the measured results. It also reduces the amount of writing during the linking of the data with the patient identification.
3. There exist modular structures laboratory programs for interactive process control. Some of these modules have to control them completely and display them adequately.

Besides these problems in the laboratory itself, in nearly all existing hospitals the patient identification data (name, prename, address, . . .) are not available in a machine readable form, so that an automatic system for patient registration must also be developed.

2. Patient reception

For patient reception we use a satellite computer (Fig. 1) with three displays for the input of the patient data plus one printer with two paper tracks for the output of all registration forms and the register cards. On a teletype we print furthermore a metallic card containing the patient data readable and the patient identification number machine-readable only.

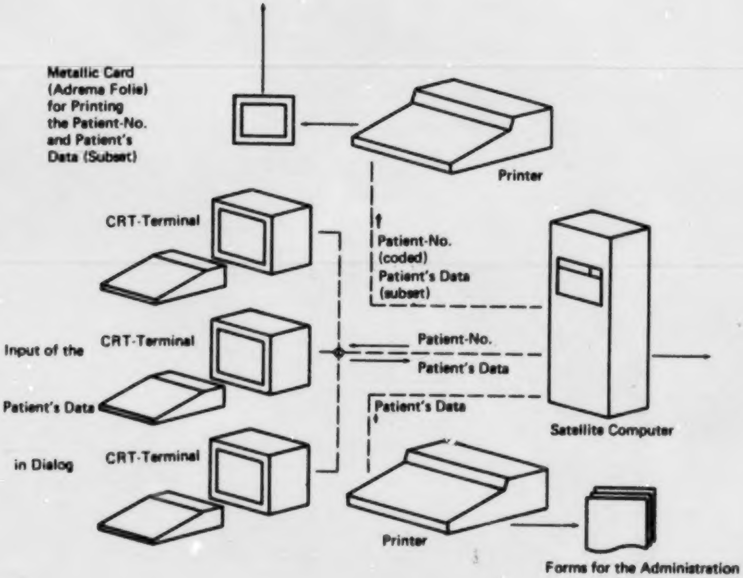

Metallic Card
(Adrema Folie)
for Printing
the Patient-No.
and Patient's
Data (Subset)
Printer
CRT-Terminal
Patient-No.
(coded)
Patient's Data
(subset)
Input of the
CRT-Terminal
Patient-No.
Patient's Data
Patient's Data
Patient's Data
Satellite Computer
in Dialog
CRT-Terminal
Printer
Forms for the Administration

Fig. 1: Patient reception system

In this dedicated computer we use no operating system because of the costs of the main memory. The controls for all the peripheral units and the datacommunication with the laboratory computer are written in machine language.

In the dedicated computer all formal controls of the entered data are carried out, and a fixed logical record is produced. To synchronize the data transfer, and to compensate short waiting times in the data transmission, 12 patient records can be stored in the small computer. Such a dedicated system has the advantage that the laboratory does not have to carry out administrative work and furthermore the system has more redundancy.

3. Laboratory test request system

Using the metallic card the nurses on the different wards can identify the laboratory request sheets in machine-readable form. The laboratory tests are ordered by marksensing sheets.

A daily number (also coded in machine-readable form) is stamped on each sheet and specimen coming into the laboratory. The sheets are then separated from the specimen and input to the computer via an on-line marksensing reader.

After centrifugation the specimen is distributed to the different working places by an automatic distribution unit controlled by the computer. This unit produces tubes which are coded with the patient identification number and filled with the material. Each instrument is thus provided with a fully identified chain of tubes. In the measuring instruments optical readers sense the patient identification and the examination number on the tubes. These identification data and the measured values are passed on-line into the computer.

4. The file system of the laboratory program system

The problem to be solved by the program system in the computer by processing the requisitions, is the sorting of a patient-number oriented mode and simultaneously of a test-number oriented mode. We use therefore two files, the patient-data file and the work-sheet file. In the patient-data file all requests are stored according to their arrival in the computer. This storage is necessary to check whether all tests requested have results at the end of the daily work. In the patient file all results are stored also.

The work sheet file has an index sequential pool-structure which allows the printing of work lists of different length by optimal use of the existing external storage capacity. The measured results are stored in a third file constructed in

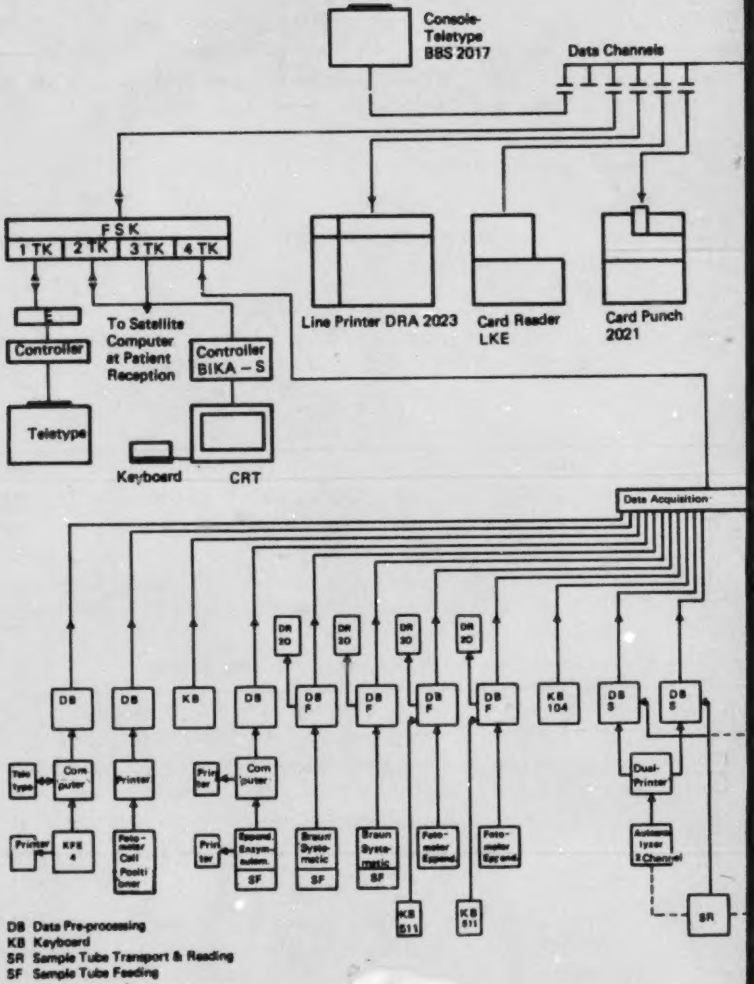
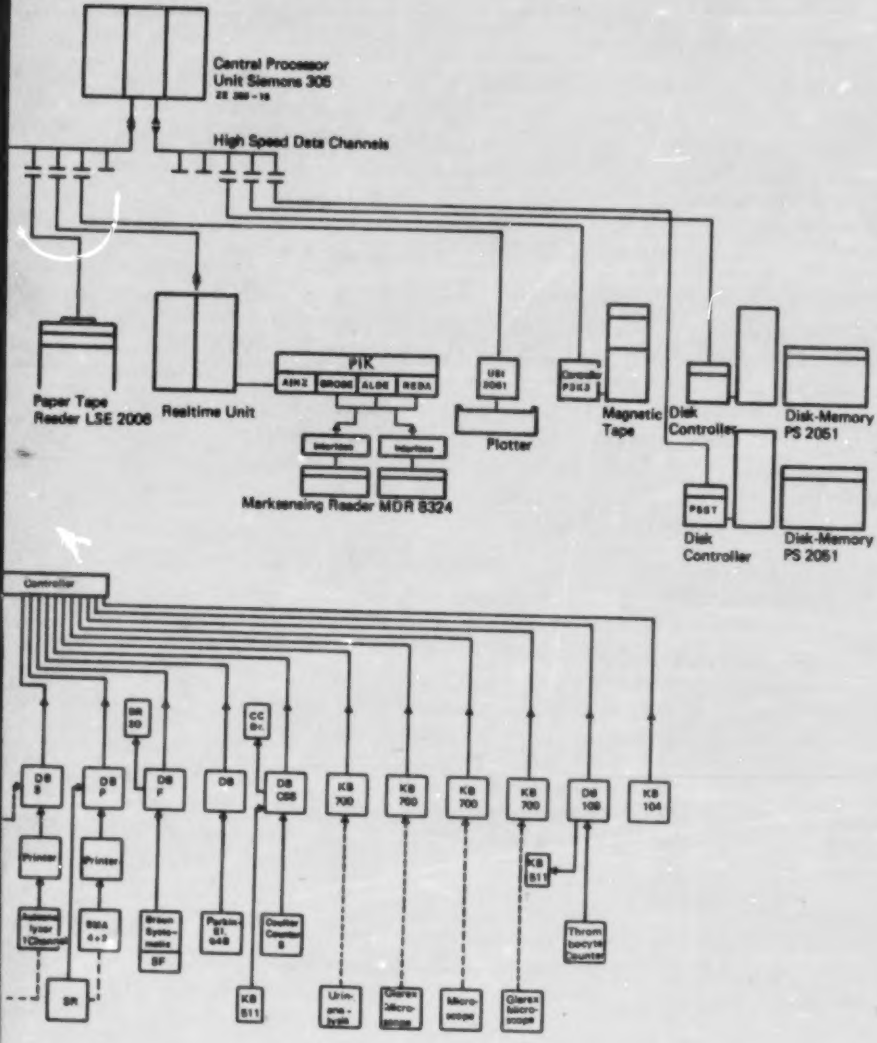

Console-
Teletype
BBS 2017
Data Channels
F S K
1 TK
2 TK
3 TK
4 TK
E
Controller
Teletype
To Satellite
Computer
at Patient
Reception
Controller
BIKA – S
Keyboard
CRT
Line Printer DRA 2023
Card Reader
LKE
Card Punch
2021
Data Acquisition
DB
KB
DB F
DR 20
KB 104
DB S
Tele type
Com puter
Printer
Prin ter
Printer
KFE 4
Foto-meter Cell Positioner
Eppend. Enzym-autom.
SF
Braun Systematic
Foto-meter Eppend.
KB 511
Dual-Printer
Autoanalyser 2 Channel
SR
DB Data Pre-processing
KB Keyboard
SR Sample Tube Transport & Reading
SF Sample Tube Feeding

Fig. 2: On line data aquisition system


Central Processor
Unit Siemens 305
2E 360 - 10
High Speed Data Channels
Paper Tape
Reader LSE 2006
Realtime Unit
PIK
AIRZ
GROBE
ALDE
REDA
Interface
Interface
Marksensing Reader MDR 8324
USI
2061
Plotter
Controller
P3K3
Magnetic
Tape
Disk
Controller
Disk-Memory
PS 2061
PSST
Disk
Controller
Disk-Memory
PS 2061
Controller
DB
S
DB
P
SR
30
DB
F
DB
CC
Dr.
DB
CSS
KB
700
KB
700
KB
700
KB
700
DB
100
KB
104
Printer
Printer
KB
511
Autoanalyser
1 Channel
SSA
4+2
Braun
Systematic
SF
Perkin-
El.
648
Coulter
Counter
S
Throm
bocyte
Counter
SR
KB
511
Urin-
ana-
lysis
Glarex
Micro-
scope
Micro-
scope
Glarex
Micro-
scope

the same way. Out of this file we print the laboratory test forms and the quality control sheets.

5. The data acquisition system

The fundamental problem in implementing real time data acquisition systems in the laboratory is the on-line connection of a great number of various items of equipment to the computer, being automatic, semiautomatic, and manual. In addition, the various commercially available analytical instruments have generally no on-line connectors nor do they provide measured curves on different analog or digital outputs. Besides the output voltage, the timing requirements for the signals and the connectors are of different kind.

The first automatic laboratory systems in Germany (universities of Tübingen and Erlangen) as well as in Sweden used computers with digital or analog processing elements to solve these problems. In consequence of this solution all command, strobe, and information signals had to be initiated by extensive software modules instead of using the core memory capacity for the calculation of results. The new systems in the University of Hannover and in our Hospital use so called pre-processing units (Fig. 2) between the computer and the analytical instruments. These units carry out various tasks, of which only the most important are listed here:

- The take up of electrically represented values and transformation of them in TTL signals
- The processing of measured values, for example peak detection, logarithmation or multiplication by a factor
- The detection of formal errors
- The linking of the patient identification and the measured value
- The display of errors detected by the program-system

The use of such pre-processing units has the great advantage that the laboratory computer receives standardized data sets from all instruments containing the test-number, the patient-identification, and the measurement result.

6. The program system

At the present state of biochemical and medical requirement, we are just beginning to realize the possibilities which data processing in laboratory can open to us. However, some of the initial problems in this aspect have been resolved in our system in München-Harlaching, now operating for three years.

- Realtime error check and display of messages for the users by a system of 4 different phases of control.
- Determination of calibration curves and precision checking also combined with messages for the user.
- Output of laboratory and patient results in a user-oriented form.
- Pattern recognition of result combinations.
- Output of a daily working strategy calculated by use of economical parameters.

7. Conclusion

For further development we want to create a series of special checks to control the chemical processes and the function of the equipment. It is also intended to develop computerized methods for a diagnostic help based on correlation of conditions. We hope that the present system has already made a decisive contribution to the resolution of these problems.

Principles of Information Retrieval Systems

Erhard Konrad, Berlin

General Principles

Three types of information retrieval systems can be distinguished:

1. Document retrieval systems which provide the user with references to documents (often used as synonymous with information retrieval systems) (6, 9).
2. Fact retrieval systems (data retrieval systems) which deliver facts and data (5, 10).
3. Inferential retrieval systems (question answering systems) which give access to implicit information (8, 10).

The development of computer-assisted information systems has led to important principles: user orientation, conceptual databases, semantic level or levels of structuredness (12), use of software tools, systemeering, data models, system models, similarity structures, feedback.

The centre of an information system is not the machine but the (conceptual) database. The system must be user oriented, that means it has to have an interface which is adequate to the mental model of the end user. The user communicates with the database and not with the machine.

In the conceptual framework of Sundgren (16) there are four semantic levels: subject matter model, infological model, datalogical model, and database. The reality-database-mapping consists of four submappings which correspond to the different models: abstraction process, specification process, design process, and implementation process.

In order to describe information structures different data models have been developed. They have different fundamental catagories. Codd (3) is starting with the concepts "elements", "domains", and "relations"; Engles (5) prefers "entities", "attributes" and "values", and Senko (14) "entities", "sets", and "names".

The techniques to build up systems are called systemeering. Langefors (17) has established a fundamental principle of systems work, that is systemeering consists of the following steps:

(1) Definition of the system as a set of parts.
(2) Definition of the system structure.

(3) Definition of the parts of the system.
(4) Definition of the properties of the system.

The fundamental principle implies that the work can be divided among different people. Different skills are required for the different tasks (1–4). In most of its tasks the principle can be mathematically formalized.

In connection with the development of information retrieval systems the following software tools are useful (Schneider (13)): text processing, searching and sorting, parsing, graph matching, classification, statistical procedures, formal description techniques, storage and access techniques, software engineering principles (modularity, adaptability, flexibility, portability), programming languages, and retrieval languages.

Similarity of information items is a practically relevant topic with application both in fact retrieval (15) and in document retrieval (10, 11). The fundamental idea is to regard the database as an abstract space where similar items are close together. Databases can be classified with respect to their completeness, scope, and density:

(1) Incomplete databases with wide scope.
(2) Complete databases with narrow scope.
(3) Complete, sparse databases.
(4) Complete, dense databases.
(5) Complete databases with wide scope, generally sparse, but locally dense.

Locally dense databases are the starting point for clustering (classification) procedures. Especially important are graphical theoretical clustering techniques which have been applied in many areas of information retrieval (2, 10). Using similarity measures different cluster types can be created: connected components, stars, and cliques e.g. The most important measures have been invented by Tanimoto (10) and Akosai (cosine measure) (10).

Document Retrieval Systems

The database of a document retrieval system consists of documents. The answer to a query is a subset of the database. A document retrieval system can be described by a vector space model (10, 6). We distinguish between three sets of objects: a set of documents D, a set of queries Q, and a set of identifiers (descriptors) I.

A database transformation T_1 maps each document into the vector space V spanned by the identifiers.

A transformation T_2 maps every query into the vector space V.

A search function S on the Cartesian product of the document space and the

query space induces a partial ordering on the set D.
A transformation T_3 on the set D generates a subset D' of D.

The retrieval process consists of the following steps:
(1) The user poses a query to the system.
(2) The query is translated into a vector.
(3) The query vector is matched against all document vectors of the database. For every query-document pair a correlation coefficient K is calculated.
(4) The document vectors are ranked in decreasing oder of K.
(5) The top n ranked documents are brought to the user's attention.

In a clustered document space every cluster is represented by a centre. The (modified) retrieval process is conducted as follows:
(1) The query is matched against the centres of all clusters.
(2) The centres being most similar to the query are examined.
(3) The documents corresponding to the most pertinent centres are individually compared.

Cleverdon (4) has established six criteria for the evaluation of document retrieval systems:
(1) Coverage (extent to which the system includes relevant material).
(2) Response time.
(3) Form of presentation of output.
(4) Effort of the user to obtain answers.
(5) Recall.
(6) Precision.

The ability of a document retrieval system to satisfy the user is called effectivity.
Recall is the proportion of relevant documents retrieved, and precision the proportion of retrieved documents that is relevant.
Systems with high precision and low recall are named OBNA systems (*O*NLY *B*UT *N*OT *A*LL), systems with low precision and high recall are named ABNO systems (*A*LL *B*UT *N*OT *O*NLY).

Excellent methods in order to increase the effectivity are connected with feedback information by the customer. Relevance feedback procedures consist in iterating a request till the user is satisfied.

There are three methods of automatic relevance feedback:
(1) Query modification (10).
(2) Document modification (11).
(3) Hybrid strategies (1).

60

The query modification consists in bringing the query closer to relevant documents (static document space), while the document modification has the effect that relevant documents come closer to the query (dynamic document space). Document modification is a radical feedback strategy, because the whole document space is adapted to the user's needs. Hybrid strategies combine query and document modification.

With respect to recall and precision document modification and some hybrid strategies give better results than query modification (1, 11).

Literature

(1) Bock, M., H.-L. Hausen, E. Konrad, H. Zuse: FAKYR – An On-Line Information Retrieval System, 1975 Conference on Information Sciences and Systems, Baltimore (Maryland), April 2–4, 1975.
(2) Bollmann, P., E. Konrad: Zur Anwendung von graphentheoretischen Cluster-Algorithmen auf das Begriffsnetz eines Wörterbuches, Conference "Algorithmen auf Graphen" of the German Chapter of ACM, Berlin, July 1975.
(3) Codd, T.: A Relational Model of Data for Large Shared Data Banks, Comm. ACM 13, 6 (1970), 377–387.
(4) Cleverdon, C.W., E.M. Keen: Factors Determining the Performance of Indexing Systems, ASLIB-Cranfield Research Project Reports, Vols. 1 and 2, Cranfield (England), 1966.
(5) Engles, R.W.: A Tutorial on Data Base Organization, Automatic Programming, Vol 7, Part 1, 1972.
(6) Kochen, M.: Principles of Information Retrieval, J. Wiley 1974.
(7) Konrad, E.: Dynamische Dokumentenräume, Lecture Notes in Economics and Mathematical Systems, 78, 499–502, Springer 1973.
(8) Konrad, E.: Deduktionsprozesse in Frage-Antwort-Systemen. In: G. Veenker (Ed.): Künstliche-Intelligenz-Forschung in der BRD, 3–14, Bonn, Febr. 1975.
(9) Lancaster, F.W.: Information Retrieval Systems, J. Wiley 1968.
(10) Salton, G.: Automatic Information Organization and Retrieval, McGraw-Hill 1968.
(11) Salton, G. (Ed.): The SMART Retrieval System, Prentice Hall 1971.
(12) Schneider, H.J.: Die Strukturiertheitshierarchie bei den verschiedenen Modellen einer Problemlösung, Workshop "Problemgegebene Datenstrukturen". In: Durchholz, Konrad, Richter, Schneider (Ed.): 4–18, 156–160, Birlinghoven, 1973.
(13) Schneider, H.J.: Bibliotheksautomatisierung und Dokumentationssysteme aus der Sicht der Informatik. In: Deutscher Dokumentartag 1973, 74–91, Verlag Dokumentation 1974.
(14) Senko, M.E., E.B. Altman, M.M. Astrahan, P.L. Fehder: Data structures and accessing in data-base systems, IBM Systems J. 12 (1), 30–93, 1973.
(15) Skinner, C.W.: A Heuristic Approach to Inductive Inference in Fact Retrieval Systems, Comm. ACM 17, 12 (1974), 707–712.
(16) Sundgren, B.: The Infological Approach to Data Bases. In: Falkenberg, Klutentreter (Ed.): Datenmodelle und Systementwürfe für Datenbanksysteme, 133–174, Summer School Birlinghoven, Sept. 1974.
(17) Langefors, B.: Theoretical Analysis of Information Systems, Auerbach Publ., Philadelphia 1973.

Retrieval Languages

Peter C. Lockemann, Karlsruhe

In developing a computerized information system for a given application area, or just in applying an already available system, one of the more difficult tasks is to establish a common ground between the intended users of the system on the one hand and the system designer on the other hand. It is suggested that the common ground may be found in the notion of "model". In order to arrive at a model of his problem area the user must clarify which aspects of his problem he deems vital or important to its solution and which ones he may safely disregard; furthermore he must explore the various relationships that exist between the relevant aspects; and finally he must describe these in abstract terms that permit a precise and unambiguous interpretation by others that communicate with him. On the other hand, the system specialist must establish abstract terms which allow him to specify those problem aspects for which he has a technical solution, or which are met by a given data base system. The task, then, reduces to the development of mutually understandable concepts for modeling, and to describe user wishes and system interfaces with their help.

Data bases are models of some real world ("universe"). Data base systems, however, usually will not restrict data bases to a single universe. Instead, because of the expenses involved they permit data bases for quite different universes (and hence users), most of them unforeseen at the time of design. On the other hand, in order to be manageable they must impose some restriction with regard to the possible structure of the models.

Correspondingly, a data base system interface is a set of tools with which to construct models (or more precisely: to communicate to the computer how to construct the model in the data base). These tools must be provided in the form of a language that allows to define the structure of the model and its components (the "data definition language"), subsequently to build up a specific model with the defined structure, modify it and delete components (the "data manipulation language"), and, above all, to retrieve components satisfying some more or less complex criteria ("retrieval language").

In order to
- describe various data base system interfaces in a consistent fashion,
- be able to compare these interfaces among themselves and with the structure of the application area,

– determine how a problem area may be expressed in terms of an interface, some general concepts for modeling are introduced. Briefly, these are the following.

(1) Modes. These are tools for abstraction that determine the general structure that the models represented by the data base must exhibit. Modes are of primary interest to the designer or manufacturer of the data base system because they still leave open the application of the system to a number of universes. Based on the notion of mode a few additional concepts are developed: Model, model object, value, elementary, and complex modes.

(2) Classes. These limit the consideration further to a single universe whose models, consequently, may be enumerated. Classes are of primary interest to the system user because he is usually concerned with a concrete, well-defined, real world. Elementary and complex classes must be distinguished.

(3) Activities and operators for the manipulation of models may be defined on modes and classes. Activities and operations differ in their degree of constructivity. Model selectors are introduced that identify the desired model objects to the activities or operators.

(4) Finally, the concept of type is introduced as a class together with a set of restrictions on the applicability of activities and selectors.

It can be shown that these concepts meet the first goal of a consistent and systematic description of data base system interfaces, and even permit the detection of some of their shortcomings. This is illustrated by three examples of modern data base system interfaces or proposals for these: IMS (Information Management System), DBTG (CODASYL Data Base Task Group), Relational model.

It is hoped that these concepts lead to a more precise description (and definition!) of the structure of the application area. In this way the frequent misunderstandings between the user on the one hand and the system designer or vendor on the other hand may be replaced by a more fruitful cooperation and, in the long run, by a higher degree of success in the application of the system.

Structures of Computerized Data Bases

Theo Lutz, Esslingen

1. Data Base Attributes

A Data Base (DB) is a
- nonredundant
- multi-usable
- independent
- physically available

set of Data Elements.

Herein means

nonredundant: Each data element is kept once and only once within the data base.

multi-usable: Whenever a data element is 'asked' for, it is made available by the data base.

independent: Data elements in the data base are independent from data representation, devices, descriptions, programs or special structures. They can be converted to any user's request.

physically available: We consider a data base not only as a mathematical term or ideal abstraction but as a physically available object.

2. The 'DB' Data Element

2.1 Real World and Data Element

The data element should describe a real world situation as shown in figure 2.1-A where *real world* is considered as being an entirety. For to describe such a real world, an *entity model* has to be formulated (8). Entities within such an entity model need to be distinct and unique, forming a set of members, as showed on figure 2.1-B. Such an entity model is able to deliver a data model where each entity model member can be represented by a data element. Relations between elements of the real world are shown as relations between entities and entity members of the entity model, then consequently interrelations between the data elements of the data model (figure 2.1-C).

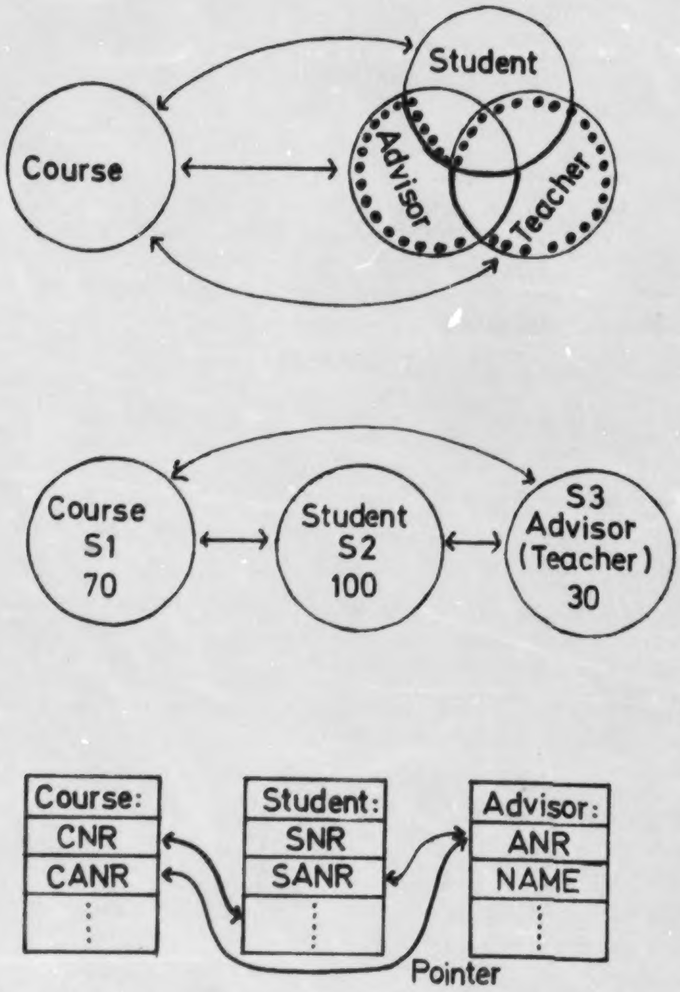

Student
Course
Advisor
Teacher
Course
S1
70
Student
S2
100
S3
Advisor
(Teacher)
30
Course:
CNR
CANR
Student:
SNR
SANR
Advisor:
ANR
NAME
Pointer

Fig. 2.1: Real world, entity model, data model

2.2 The Role of the Data Element

Data elements from a data base can be grouped according to user requirements. Such a group is called a *logical record*. By *physical record* the data element within the physical available data base is meant.

The data element itself is a vehicle which carries information from the physical level to the logical level during data base queries and it carries information from the logical user level to the phsical level during updating. This influences the size and partitioning of the data element. If data elements are fragmented too much then too many are needed to built up one logical user record (access redundancy). If data elements are too big un-needed information is carried to the logical level and vice versa.

To find an optimal partitioning of data elements is one of the first steps during data base design.

2.3 Types of Data Elements

A data element is called to be

formatted if specific fields can be inserted into a predefined scheme called the format of the data element

unformatted if such a scheme does not exist, but general rules like a grammar (syntax), where common words of a common language are grouped forming data elements with no predefined format.

Uncoded if no discrete codification is applied or available. Then the data element is represented by a general picture or symbol.

All further considerations within this paper are concerned with **formatted data elements only**.

2.4 What Element means

As the term element means something like *atom* or unit a proper definition among appropriate data processing units is needed.

Technical data processing "atoms" are:
  Bit, Byte, Word, Segment, Record, File
Logical data processing "atoms" are:
  Character, Field, Domain, Data Element, Set of Elements

As the data base view mainly is a view to the structuring and retrieval of data, fields or words can not be the data base atoms, although fields are used for data base evaluation. However a grouping of fields seems capable of forming the data base data element.

A group of characters is called a field or a domain. *A group of domains being kept in a self-contained way is called a data base data element! It is the atom of the data base with respect to structural and retrieval problems.*

2.5 The Formatted Data Element

A formatted data element is now a data base data element (as a group of domains), where the sequence of domains as well as domain attributes are defined in terms of a format. The *format* is a predefined scheme which defines the structure of the data element.

Domains used for identification of individual data elements in a data base are called *key domains*. All other fields are called *value domains*. If a formatted data element has n keys and m values its format can be described as

$$DE_j = DE_j (K_1 \dots K_n ; V_1 \dots V_m)$$

Let j vary from 1 to N, all N different types of formatted data elements in our data base ar nominated. j is called the *name* of a data element. It stands for a set of data base elements being all of the same type or structure. If the data base is really nonredundant the N sets of data elements in a data base are disjoint which says that they do not have data elements in common.

Keys are used to *select* data from one set of data elements and to *synchronize* data elements from different sets of data elements.

Nomination, selection, synchronization, and evaluation describe the basic operations on a data base.

2.6 Data Element Classification (Access)

The occurence of keys in a data element of a specific type can be used to distinguish between data elements.

A data element is called *simple* if it has one and only one key available.

$$DE_{simple} = D_E (K, V_1, \dots, V_m)$$

A data element is called *complex* or *multiple* if it has more than one predefined key.

$$DE_{complex} = DE\,(K_1 \dots K_n; V_1 \dots V_m)$$

If no key domain is predifined and if the property being a key can be assigned ad hoc to any element's domain, the data element is said to be *absolute*.

$$DE_{absolute} = DE\,(F_1 \dots F_k)$$

The data element attributes simple, complex, or absolute are mainly attributes of appropriate data base processors.

2.7 Data Element Classification (Structure)

A data element is called *simple structured* if it does not contain repetitive groups of domains especially if the repetition of groups is performed with a variable number of elements. Within the sequence of domains, e.g.

Father's Name – Children's Name

the domain *children's name* will be repeated for each child.

A data element is callen *complex structured* if it comprises repetitive groups of domains.

Sometimes data elements being simple structures but having a self-contained nature by the presence of keys are grouped together. Such a data element is designated as being of *compound structure* (Figure 2.7).

We will be concerned below with simple structured data elements, but our main concern will be with complex data elements.

3. Relational Description

3.1 Relational Basics

Let $D_1 \dots D_N$ be N distinct sets where d_n is an element of the set D_n. The D_n are called *domains* or *domain sets*.

If we now built a new set of the type

$$R_m = R_m\,(d_1, d_k \dots d_i)$$

where d_1 is taken from D_1 and so on, but not necessarily all combinations potentially possible are taken, R_m is said to be a *relation* on the domains D_n. R_m is obviously a subset of all potential combinations on the domains following the above pattern. This complete set is called the *cartesian product* of the domains involved.

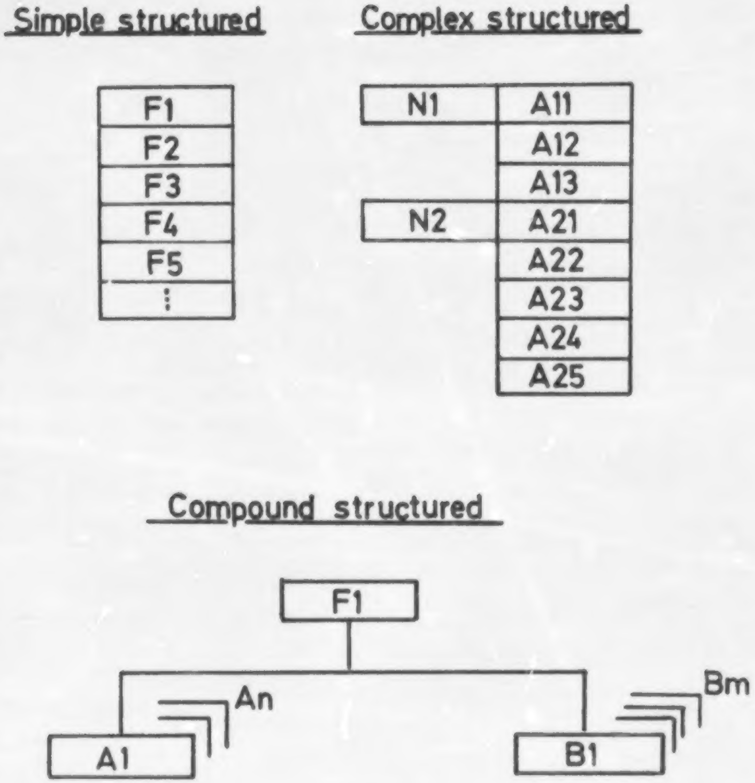

Simple structured
F1
F2
F3
F4
F5
⋮
Complex structured
N1
A11
A12
A13
N2
A21
A22
A23
A24
A25
Compound structured
F1
An
A1
Bm
B1

Fig. 2.7: Data elements structures (simple, complex, compound)

The definition of a relation, which can be found in all books describing set theory, is very close to our above definition of the data base data element. Consequently the term 'relation' can be used to describe the data base data element. Each relation is evidently free of redundancy as the definition of a set excludes the occurence of duplicates. If now $R_1 \ldots R_N$ are the N relations identical with what we previously have defined as types of data elements, the general union of those relations defines the data base. To suppress further redundancy we now ask for the involved relations being disjoint. Figure 3.1 gives a demonstration of the terms used.

3.2 The Role of Keys

The use of domains in a relation gives a better understanding of the role of keys in a data base. If D_k is a domain in a relation whose elements can be used to identify individual tupels of the relation, the D_k is said to be a *key* of the relation.

If the elements of D_k occur uniquely and nonredundant within the relation, D_k is called a *main key* or *primary key*. Else D_k is a *secondary key*. Further it could happen that a main key is used to establish a specific order among the tupels. Such a main key is called a *master key*.

One needs to be aware that a domain D_k can be used as the main key or even the master key in a relation but being only a secondary key or no key in another relation of the same data base.

Also it may happen that a relation has no unique domain. If for example one builds up a relation showing students having enrolled for courses where one tupel is available if a certain student goes to a certain course the students' number as well as the course number will be redundantly used within such a relation as one student usually goes to various courses and one course is taken by various students. If now a master key is needed, uniqueness is only achieved if the student's number domain is concatenated to the course number domain being now a new domain which is unique.

On the other hand it never will occur in a well defined relation that there is no key available for as a relation is a set and as a set is an entirety of well distinguished elements at least the concatenation of all domains must lead to uniqueness as no duplication is kept in a set. Consequently a data base relation has at least one unique key and has consequently always a master key.

Format
1 CNR
2 DESCR
3 TEACHER
4 ROOM
5 LOAD

Format

E_1	E_2	E_3	E_4
110	112	115	119
A	C	B	D
166	169	166	166
37	39	32	39
4	2	2	1

Individuals

Domain Sets
{110, 112, 115, 119}
{A, B, C, D}
{166, 169}
{32, 37, 39}
(1, 2, 4)

Domain Sets

DB-Domains $D_n \quad n=1.....N$

DB-Relations (sets of tupels)

$$R_m = R_m(D_\ell, D_k, \ldots D_i)$$

Non-Redundancy

$$R_i \cap R_j = \emptyset \quad i \neq j$$

Data Base Definition

$$DB = \bigcup_{1}^{M} R_m$$

Fig. 3.1: Relational DB-description

3.3 Normalized Data Base Relations

If we understand our data base data elements as elements of relations it could happen that one of the domains used to build up a relation is itself a relation. By this we do not mean that, for example, a student number domain is a relation of the characters or digits used to build up the individual domain elements.

But if we have a relation for each advisor, showing the courses he helds within a domain *course*, this domain could be a relation where elements occur in a variable number for each advisor. This is exactly the same case as where a data element comprises a repetitive group of data. The following formal structure can make this a little bit more understandable.

Let

ADVI (ANR, A1, A2, A3, *course*)

be a relation having one element for each advisor so that ANR is a unique domain. Further let

course (CNR, C1, C2)

be the relation which is used within the ADVI relation. This was previously called a complex structured data element.

If we enhance now CNR by a left attached ANR to ANR-CNR using for ANR the value occuring in the tupel in which *course* occurs, the relation course is unique with respect to ANR-CNR, as well as the interrelation between the advisor tupels and their courses is unique. So if we delete *course* within ADVI and if we take the enhanced *course* definition we no longer have a relation within a relation. Our data elements are now simple structured. The process of going from the complex structure to the simple one is called *normalization to the first normal form*. Simple structured data elements are said to be in *first normal form (1NF)* which we assume for our data base elements. Further normalizations for data (2NF, 3NF, Index Normalized Form) as well as examples for normalizations are given in literature (1, 2, 9).

3.4 Final Remark

The use of relations to define data elements of a data base, as well as for the definition of the data base itself, gives a very proper and mathematical understanding of the data base problems, even if this version of understanding data bases is at a relatively early stage. The relational model for data bases has rised so many serious questions that without any doubt further data base research and development needs to use this tool. Special attention is directed to the problem of non-algorithmic descriptions of data base use through special relational operators (see 1, 5).

4. Pointers, Keys, Structures

4.1 Keys and Pointers

If a logical relation goes from one data element to another data element, a *pointer* is said to be applied. We call it an *abstract pointer,* as long as this is only a logical pointing relation between data elements. Vice versa, if such a pointing relation, which is in many ways a directed pointer, is physically realized by storing within the data base an interrelation between a key and a data element location, such a *concrete pointer* is called an *index* or an *index pointer*.

The data element location, say on a direct access, storage device, is usually expressed in terms of software or hardware addresses. By using a key as entry to the system a data elements address is achieved and consequently the data element itself.

A pointer, pointing to no data element is called a *null pointer*. This type of pointer is needed to achieve completeness in representation.

Attributes of pointers could be:

Singleton: This pointer points to one and only one data element.

Twin: A twin pointer points to a data element of the same type on the same level.

Child: This type of pointer goes as being directed to another type of data element in the data base and at a lower level.

Parent: This is the inverse pointer of a child, by going in level up to another type of data element.

Forward is the inversion of a *backward pointer*, forward and backward showing a preferred direction, which is not expressed in terms of singelton or twin.

Further a concrete pointer may reside inside or outside a data element where it is pointing from (home element). In case the pointer does reside inside and is not a singleton, it is part of an *address chain* or of a *multilist*. In case the pointer resides outside is part of an *indextable* or an *inverted file*. Address chains and index tables are interchangeable techniques, as it is only depending on the fact where the pointer is residing.

Another difference in pointers could come from the key underlayed. A secondary pointer applies to a secondary key, a primary pointer has a main key as a base. If for example secondary pointers are applied by pointing to some domain values each various address chains (multi lists) go through the data base. For each domain value we would obtain a singelton pointing to the first data element with the assumed domain value. Then the chain will be continued by twins until a null pointer marks the end of the chain. The singelton in such a system, being the first pointer on the chain, is sometimes called an *anchor*. Anchors can be kept in an index table, which is a special index called the chain directory, as it defines the entry to the set of data elements. Obviously anchors can be stored inside data elements. Even the examples used show, that a variety of techniques can be applied in using concrete pointers, which really can be found in present systems; up to now we did not use any technical but only logical reasoning. Figure 4.1 shows how concrete pointers can be applied to form an index based data base with a narrow starting from a box being used as a graphical symbol. It also shows the role of null pointers.

A data base built on concrete pointers is said to be physically structured!

A data base built on abstract pointers is said to be logically structured!

4.2 Tree Structures

Figure 4.2 shows a very common physical structure for data bases. Available is a set of data elements called the roots. They may be interconnected by a primary chain following the master key. Child pointers go from each root to another level which forms the first level within the tree, and further children go from each non-root level data element to deeper levels, until null pointers mark the end. Data elements within such a system having null pointers are called terminals, but terminals may occur on any level.

The fact that each data element has a well defined level, being of well defined type, and that each data element is pointed at once and only once (indegree 1), and that each data element points to various data elements (outdegree n) opens a unique way from each terminal to its root. Such a unique chain is called a path or a branch and the entire structure can be understood as an overlay of such branches.

Within each box of our structure being called a tree structure or a hierarchical structure, we have marked by a capital the name (type) of the data element. On its right the level of the data element is indicated. Whenever more than one data element of the same type occurs on the same level, the occurence of that type is counted and the tally's value is shown.

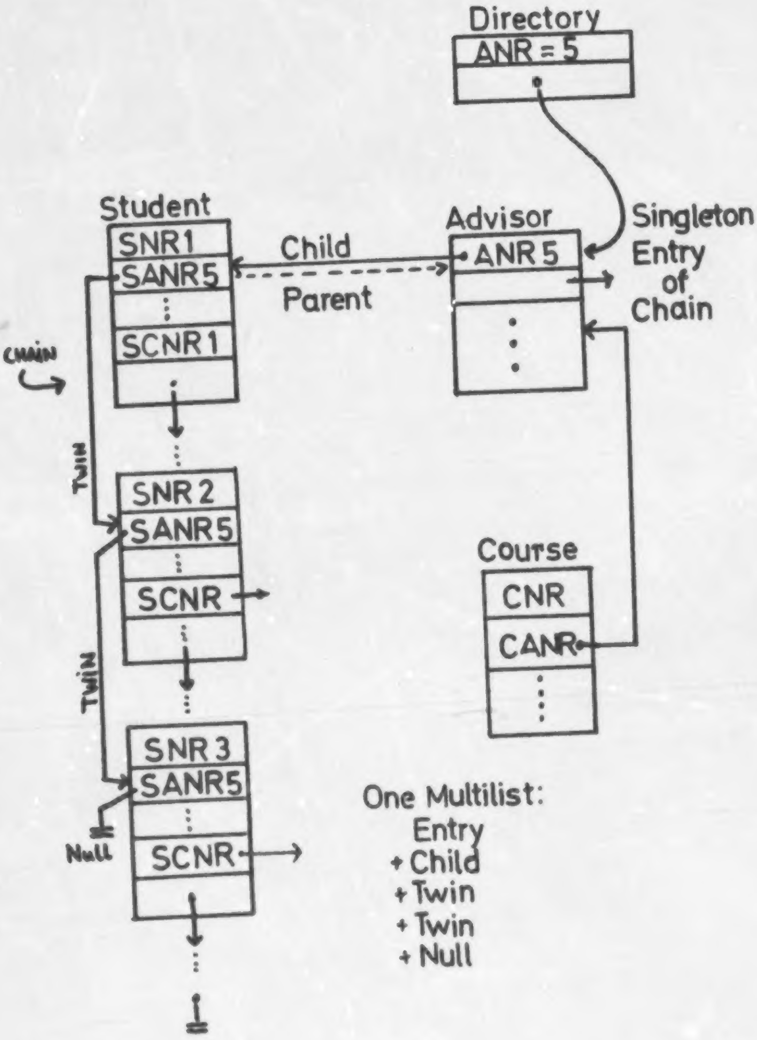

Directory
ANR = 5
Student
SNR1
SANR5
SCNR1
Advisor
ANR 5
Child
Parent
Singleton
Entry
of
Chain
CHAIN
Twin
SNR 2
SANR5
SCNR
Twin
Course
CNR
CANR
SNR 3
SANR5
SCNR
Null
One Multilist:
Entry
+ Child
+ Twin
+ Twin
+ Null

Fig. 4.1: Keys and pointers in a data base

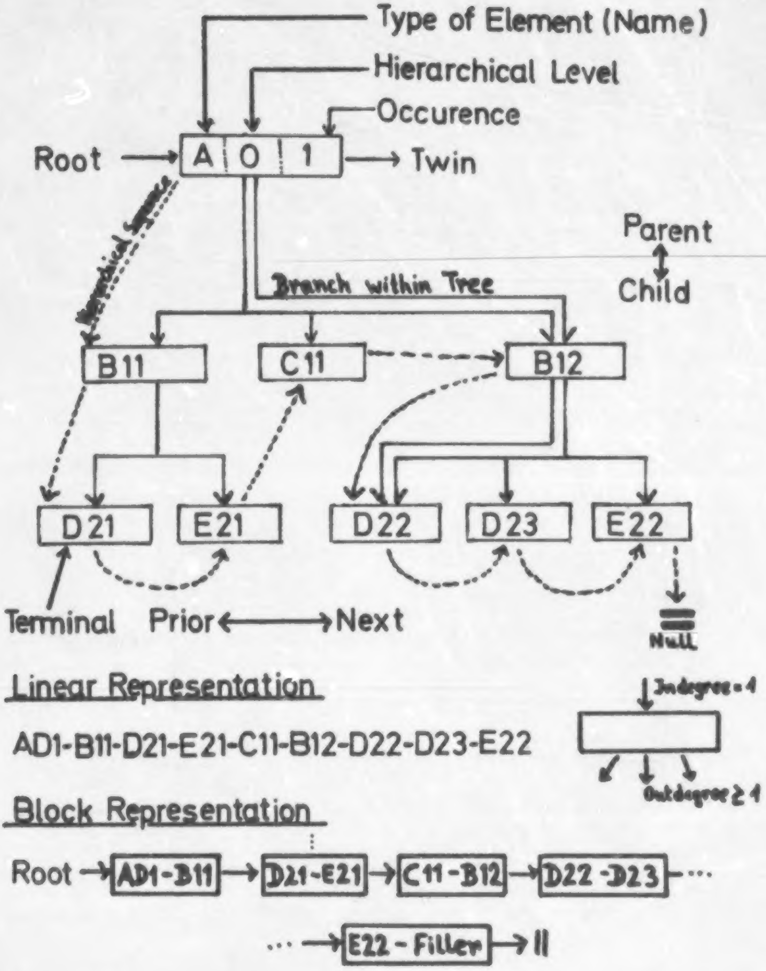

Type of Element (Name)
Hierarchical Level
Occurence
Root
A
0
1
Twin
Hierarchical Sequence
Parent
Child
Branch within Tree
B11
C11
B12
D21
E21
D22
D23
E22
Terminal
Prior
Next
Null
Linear Representation
AD1-B11-D21-E21-C11-B12-D22-D23-E22
Indegree = 1
Outdegree ≥ 1
Block Representation
Root
AD1-B11
D21-E21
C11-B12
D22-D23
E22-Filler

Fig. 4.2: Tree structure

If we now apply a convention that starting at the root we go from top to bottom until we reach a terminal, and from left to right following the occurence we have rules to arrange our tree structured data elements in a line which is a unique order.

Due to the line structure this representation of our tree structure is called the linear form of the tree.

For technical reasons – for example within an agreed byte length (block length) – data elements are grouped to blocks which then can be kept as storage units on storage devices, and then interrelated by block pointers realizing the linear order of the data elements.

Due to the variable occurrence of data elements of the same type, the length of the linear form varies from root to root. By introducing the block representation of the tree only a variable number of blocks remains, which is technically easier to maintain through appropriate block pointers than the variable length of the line. The fact, that the tree structure due to its relation to the linear form, can be kept on linear storage devices makes the tree important for all storage considerations.

4.3 Groups of Data Elements

If we write close to the IBM-Query Language GIS (Generalized Information System) the user request for data in an algorithmic way:

```
QUERY ADVISOR, STUDENTS, COURSE
FOR EACH ADVISOR
WHEN ANR EQ 166
LIST ADVISOR
   FOR EACH STUDENT
   WHEN ANR: STUDENT EQ ANR: ADVISOR
   LIST STUDENT
   AFTER LAST STUDENT
   FOR EACH COURSE
   WHEN ANR: COURSE EQ ANR: ADVISOR
   LIST COURSE
   AFTER LAST COURSE
AFTER LAST ADVISOR
END: QUERY
```

we will obtain as first data element the advisor element we assume to be available. Next all student elements having the same advisor number as the advisor will occur by exhausting the entire data base for data elements meeting the appropriate condition. Those will be followed by course elements meeting the other condition. The result of the query is at first a line of data elements.

If we assume we would not know more about the student and the data element than that they comprise the ANR = 166, we have no reason to sort them. Thus it is quite reasonable to represent the structure of the group of data elements found by the above query; not by a line but by a tree, whose root is *advisor* having as children the terminals *student* and *course*. If from students and course on we would ask for other types of data elements, the appropriate tree would have to go deeper and would be spread out broader.

In any way the tree structure seems to have importance not only for the physical view of data, but also for the logical view, as our query represents one way to use a data base out of many in a multi-used data base.

One should be aware that linearity does not come only from the way we wrote our query, but is a result of the storage devices being linear, as they can only keep one byte after the other.

4.4 Physical Data Base Structure

Figure 4.4 shows a file sequence of data elements (FO = File Organization) and three other logical records grouped as lines where we have shown how a tree can be linearized and how the line becomes a tree.

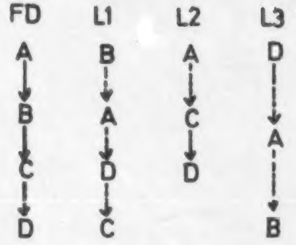
Obviously the structure of our data base must be such that it can deliver those 4 structures without having any redundancy on the physical level. This leads to the logical data base structure being an overlay of the structure of the logical user records. We realize immediately, that we now have nodes in our graph (Fig. 4.4) having an indegree greater than 1 and having outdegree greater than one.

This fact does not permit us to apply the above conventions to go from top to bottom as there is no preferred root, and to go from left to right as there is no such order. This means that one can not transform the data base structure to a mere line. Consequently the data base structure as an overlay of hierarchically structured trees is no longer of linear order.

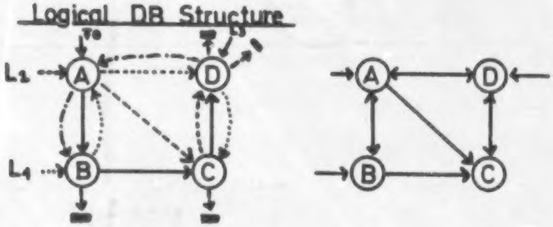
A nonredundant, multi-used and physically available data base is equivalent to a set of data elements with a non linear type of order.

This result is considered the most effective criterion for a data base especially versus conventional files. Figure 4.4 shows how such a nonlinear structure

Logical Record Level


FD
A
B
C
D
L1
B
A
D
C
L2
A
C
D
L3
D
A
B

Logical DB Structure

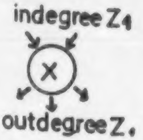

Fo
L3
L2
L4
A
B
C
D

overlay

simplified

Physical DB - Structure

(DB-Organisation)


indegree Z4
X
outdegree Z.

Location	799	800	801	802	803
Data	Directory	A	B	C	D
Fo	800	801*	802	803	Null
L1	801	803	800*	Null	802
L2	800	802*	●	803	Null
L3	803	801	Null	●	800*

* Root

● Not used

Fig. 4.4: Data base structure

which is called a general network or a topographical structure can be realized by adding a directory to the file, and by enhancing each record with a system of concrete pointers. Obviously overheads occur if a data base is physically structured. A logically structured data base achieves its structure by the structure of queries, reflecting the linear structure of user data groups (10).

5. Data Base Organization – Physical DB View

5.1 The Physically Unstructured Data Base

Such a data base organization can be obtained by having a file processing system available that can transform user queries to appropriate groups of data elements. The most trivial way to realize a simple query as

```
QUERY
FOR EACH STUDENT
WHEN ANR EQ 166
LIST ADVISOR
AFTER LAST STUDENT
END QUERY
```

is to scan through the entire file and to check for each data element whether the used keys fit to the condition applied. Complex queries occur, if simple queries are nested within other queries, or concatenated to those. Even then file scanning can be applied, however in case of synchronizing conditions (WHEN ANR IN STUDENT EQ ANR IN ADVISOR) the secondary file will be scanned for each data element from the first. This can lead to intolerable performance. Advantages of file scanning are:

+ Data base easy to maintain
+ Each domain can acts as key
+ Minimal overhead
+ Normal files (OS) as data base

Disadvantages are:

– Need for temporary files
– Need for *resort* or temporary files
– Limited efficiency
– Limited number of files involved

Performance of file scanning mainly depends upon the cleverness of the file scanning processor.

5.2 The Hierarchically Structured Data Base

5.2.1 The Pointer-free Organization

Some systems keep the logical group of data elements most used physically on the device by having *one* tree structure stored. Further user requirements are satisfied by scanning through this organisation. One example for this philosophy is IBM's system IMS/I the previous version of the actual IMS (IMS = Information Management System).

One tree structure is preferred, and appropriate trees are stored in blocks derived from the line form of the tree. The first block containing the root of the tree is kept in an index organized environment and a primary index table allows access to the main key of the root. Block pointers point to the following blocks being kept in an overflow area. This physical organisation of a data base is called HISAM (Hierarchical Index Sequential Access Method). File scanning through this system is facilitated by the fact that numerous data elements are accessable through one disk access to a block.

5.2.2 Use of Pointers

If each data element is enhanced by a control area (prefix) that can cover pointers, and if it has a unique address being obviously composed by the block address and the position within the block, then this system has a physical network structure as each data element can have an indegree greater than 1 and can have an outdegree greater than 1. Various pointers can go to one and the same data element, which itself has various pointers leaving it. Such an organization is used by IMS/II being the base of the actual IMS/VS version. The data organisation is called HDAM (Hierarchical Direct Access Method). H indicates that the data elements are physically represented by one preferred tree structure.

5.3 The Physical Tree Overlayed Network

The following relation, sometimes called a many-to-many relation, is taken from the manufacturing application. It shows in its first domain a "super-part" number, in its second domain a "sub-part" number and finally a quantity by which the sub-part occurs within the super-part. This represents a relation on the same class of objects, namely product parts, since parts are related to parts.

SUPER	SUB	QUANTITY
A	B	2
A	D	2
B	1	1
B	2	2
C	E	4
C	4	2
D	B	1
D	E	1
D	1	2
E	2	1
E	3	3

The binary relation based on the part numbers within this relation can be used to design a graph showed in figure 5.3-A. Obviously this graph is of a nonlinear type because nodes have indegrees as well as outdegrees being greater than 1. Also this graph is a nonredundant representation of the above relation because no node is repeated within it. Figure 5.3.-B shows a decomposition of the graph to binary trees; "binary" because only two levels apply. Finally figure 5.3-C shows two trees derived from the binary trees with redundancy in nodes taken into account.

The competition between trees and general networks leading to extensive discussions in recent literature has redundancy as a pivot, which is important to see. It seems, that this competition has no serious background, as networks can be avoided by taking redundancy into account and *vice versa*. Figure 5.3-D shows the representation of a pointer system, according to above relation, as a physically available and nonredundant data base. This idea is realized in the DBOMB-principle (DBOMP = Data Base Organization and Maintenance Processor) used on IBM computers.

5.4 Final Remark

The physical organization of a data base can be done either with a logically structured data base and a physically structured data base or with a physically structured data base with a hierarchical structure and a physically structured data base having for each type of data element a file for itself. There is also a range of possibilities in between for further methodological approaches. Theoretically it is very difficult to evaluate the principles. One should be aware that technical realizations of those principles open a wide field of possibilities to complicate the understanding of data bases! This underlines, why data base *design* is a very important phase in building up a data base.

Fig. 5.3-A: Network structure

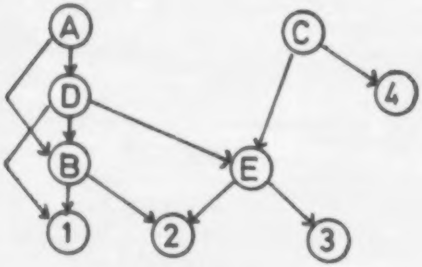


Fig. 5.3-B: Graph decomposition

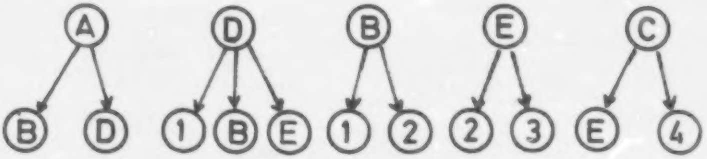
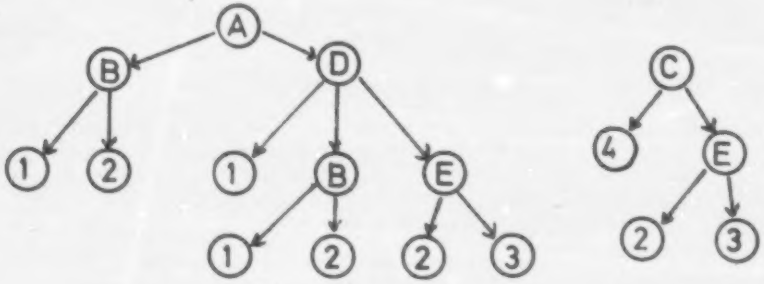

A
D
B
E
C
B
D
1
B
E
1
2
2
3
E
4

Fig. 5.3-C: Multi-level tree


A
B
D
C
1
2
1
B
E
4
E
1
2
2
3
2
3

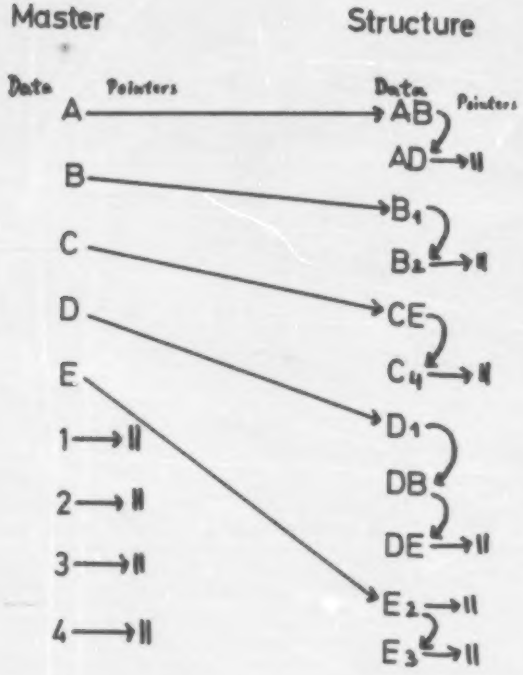

Master
Structure
Data
Pointers
A
B
C
D
E
1
2
3
4
Data
Pointers
AB
AD
B1
B2
CE
C4
D1
DB
DE
E2
E3

Fig. 5.3-D: Physical organization of a network by pointers

6. Data Base Access – Logical DB View

6.1 Data Independency through Data Description Tables

Data description tables are a very important tool to achieve data independency within a data base. Conventionally the description of the data elements and of the attributes of their domains has to be done in each program using such a data element this leads to errors and misunderstanding and altering problems. It is adequate to the idea of a multi-used data base not to leave data description to the user, but to add it to the data base data. Reference to the data used is always made by using the name of the data field. Technical attributes are added automatically when a query enters the system during compilation time of a query. So data description tables give a high degree of data independency for the user, as he can use data according to his wishes. Adaptation to the user is done by using the original data description table comprising the attributes of the physical version and the users' request for changing attributes, if asked for.

Data Description Tables (DDT) are part of most data base systems.

6.2 The Host Language Approach

A quite reasonable approach for the access of data bases is to pack all data base access functions into a data base processor system and to make this system available for various commonly used programming languages acting as host languages for the processor. Historically this was the first way and is still used, mainly because it gives a high degree of language independency. So data base processors are available for COBOL, PL/1, FORTRAN or ASSEMBLER. All access functions are described in macros bridging to the data base processor and describing mainly the retrieval and updating functions. Evaluation and all other processing is done through the notation of the common programming language, said to be the host language. Two types of macros are basically available. One type, being of descriptive nature, shifts all parameters to the processor. Another type being imperative describes how data elements have to be accessed and retrieved. A typical descriptive macro for the IMS-DL/1 data base macro language is the "Segment Search Argument" which describes in terms of conditions which data elements have to be retrieved. *Get unique* is of imperative type and asks for a well described or first data element to be accessed. *Get next* asks for a successor within a logical chain of data. *Insert* gives data elements to the data base. The same language is applied to messages entering and leaving the system.

The macro *get next* raises a typical data base problem as it could easily happen that no next element is available. Such information is communicated through a status indicator to the host language, which can take appropriate action having analyzed the situation. Signals for data base exhaust or even errors (e.g. somebody asks for data elements not defined) are of the same type of information. The following small programming outline builds up an outgoing message caused by an incoming message showing an advisor whose number is coming through the message of his data and his students:

```
      OPEN
      GET UNIQUE MESSAGE–IN
      MOVE ANR: MESSAGE TO ANR DATABASE
      GET UNIQUE ADVISOR, SSA :ADV
      IF STATUS EQ ERROR GO TO . . .
      MOVE ANR: ADVISOR TO ANR: SSA
  A:  GET NEXT STUDENTS, SSA: STUDENT
      IF STATUS EQ END GO TO END
      LIST, MOVE, COMPUTE MESSAGE
      GO TO A
END:  INSERT MESSAGE
```

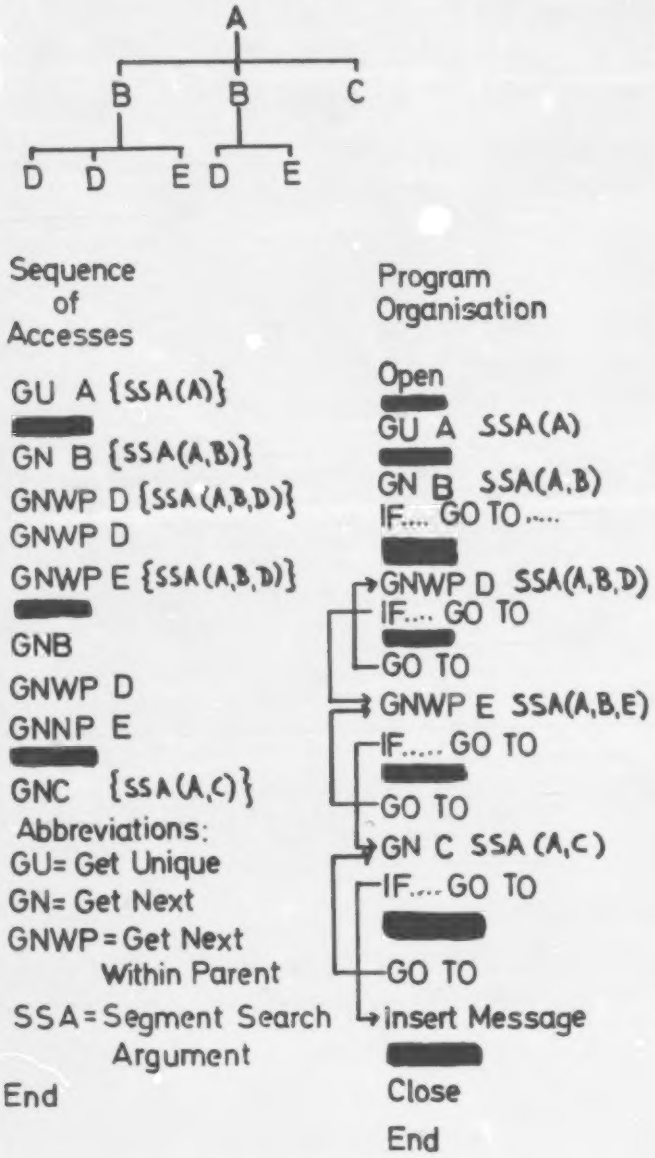
COBOL is assumed as host language and SSA is the abbreviation for a "Segment Search Argument" as a descriptive macro which must be available for each type of data element retrieved. Figure 6.2 shows a logical user record and the sequence of accesses needed to build it up. This sequence of macros is achieved by organized macros forming a program that reflects the tree structure of the user tree. Data element types are nominated by capital letters.

6.3 The Selfcontained System Approach

Quite another direction of solving the data base access problem is to develop a selfcontained system with its own high leveled language called a *query language*. GIS being previously used for demonstrating major data base problems is *such an approach for the data base access mainly realizing file scanning for accessing the data base*. The present state of the art is exclusively described by the term of algorithmic description of data base access even in the approach of selfcontained systems. The relational model for data bases opens a field of research and development which could deliver in the distant future better language approaches than algorithmic ones (see part 8).

6.4 The Language Enhancement Approach

Quite another approach occurs when starting to enhance available programming languages for data retrieval purposes by defining appropriate keywords and language facilities. This was done by a CODASYL Task Force which has enhanced actual standardized COBOL (ANS COBOL). As this approach is not yet definitively finished it is recommended that the reader reads for himself the appropriate task force report (11).



Sequence of Accesses

```
GU  A {SSA(A)}
[redacted]
GN  B {SSA(A,B)}
GNWP D {SSA(A,B,D)}
GNWP D
GNWP E {SSA(A,B,D)}
[redacted]
GNB
GNWP  D
GNNP  E
[redacted]
GNC   {SSA(A,C)}
```

Abbreviations:
GU= Get Unique
GN= Get Next
GNWP = Get Next
Within Parent
SSA = Segment Search
Argument
End

Program Organisation

```
Open
[redacted]
GU  A   SSA(A)
[redacted]
GN  B   SSA(A,B)
IF.... GO TO .....
[redacted]
GNWP D  SSA(A,B,D)
IF.... GO TO
[redacted]
GO TO
GNWP E  SSA(A,B,E)
IF..... GO TO
[redacted]
GO TO
GN C SSA (A,C)
IF.... GO TO
[redacted]
GO TO
Insert Message
[redacted]
Close
End
```

Fig. 6.2: Host language programming

87

7. DB Systems – Technical View

7.1 Query Language Modules

The host language approach for a data base access system demonstrates that the mere retrieval problem can be developed as a selfcontained module, which can afterwards be embedded in different processing environments. Other modules can be added and the following collection of modules shows the range of a query language.

7.1.1 Describe-Define

Data Base Description Tables (DDT) are kept within the data base. This module puts the tables in provided facilities to print DDT for user purposes and to maintain them.

7.1.2 Retrieval

The retrieval module comprises macros for accessing data elements and combining data elements of different types (selection and synchronization).

7.1.3 Hold/Keep

This module provides facilities to keep selected data base data within temporary files during one procedure *(hold)* or over various procedures for various users *(keep)*.

7.1.4 Edit/Layout

Statements are defined for listing data elements in a standardized way *(list)* or in a user defined way *(report)*. Possibilities for arranging and rearranging data are provided.

7.1.5 Compute

Computational facilities are requested

Count, total: fieldwise computation through the entire data base

When: defines conditions for data element selection (segmentwise computation) + – / x for fieldwise computation

7.1.6 Maintain

This module updates and maintains the data base on various levels:

Modify – to maintain data base fields
Update – to maintain a segment level
Create – to maintain a data base file level.

7.1.7 IS-Interface

Interface is needed to the information system where the data base is embedded. Facilities are logging, transmitting jobs and data to the information systems facilities for security and break down recovery and so on.

7.2 DB/DC System Components

Figure 7.2 shows technical elements which built up a data base/data communication (DB/DC) system. Those are:

+ *Data Base Files* for stored data elements comprising a logfile and message queues
+ *Physical File Relations* realizing concrete pointers within the data base
+ *Data Base Description* provides parameters and DDT
+ *Physical Data Base Access* are hardware oriented standardized channel programmes (access methods)
+ *Data Base Access Statements* are user (software) oriented access statements and descriptive macros
+ *High Level Language* as host language or specialized data base query language
+ *Procedure Bank* is the instrument which keeps predefined data base transactions available for endusers
+ *Message Control* opens the area of data communication facilities for message interface between user and system resources
+ *Terminal Control* is a selfevident element as well as
+ *Line Control*
+ *Supervisor* is an enhancement of the standardized operating systems supervisor for DB/DC
+ *Utilities* are needed to support the operating system as well.

Those elements permit the definition of different types of DB/DC-Systems. If for example no procedure bank is available, but a compiler for a high level query language is an *absolute system* is defined that permits unrestricted access to the data base data elements within the range of the language options.

If no such compiler is available, but a procedure bank which comprises all needed transactions is, a *specified information system* is defined that can deliver all information defined by the transactions stored within the procedure bank. GIS (Generalized Information System) for IBM Systems is an example for a DB/DC-System of the first type, IMS (Information Management System) for the second type.

The extremely high variety of technical details to be considered if DB/DC-techniques are discussed does not permit any systematical approach within this paper. The reader interested in this area may start by reading some of the introductory manuals for existing systems, such as those cited or others from other manufacturers.

8. An Outlook

The topics discussed in this paper are always under development and research. Especially the relational description of data base problems has opened a wide field of research problems with a strong application of mathematics. Data base research surely will achieve interesting results during the next years. Even if data base research will not immediately result in improved software systems it may give a good firm base to understand and to handle DB/DC-problems.

As we previously have discussed only the relational description of the data base problem, it might be of interest to enhance this picture by relational operators and predicate calculus items to show the reader which direction the relational research will probably go.

Let A be a relation describing *advisor*, and S a relation for *student* with the following tupels:

A:

ANR	ANAME	ROOM	FAC
166	A	4	1
167	B	3	2
169	D	1	2
181	C	5	1

S:

SNR	ANR	NAME
210	166	a
280	169	c
294	166	b
298	181	c

As relations are sets we can apply on those sets the normal set operations as *union, intersection* or *difference* to derive other sets from the above. More

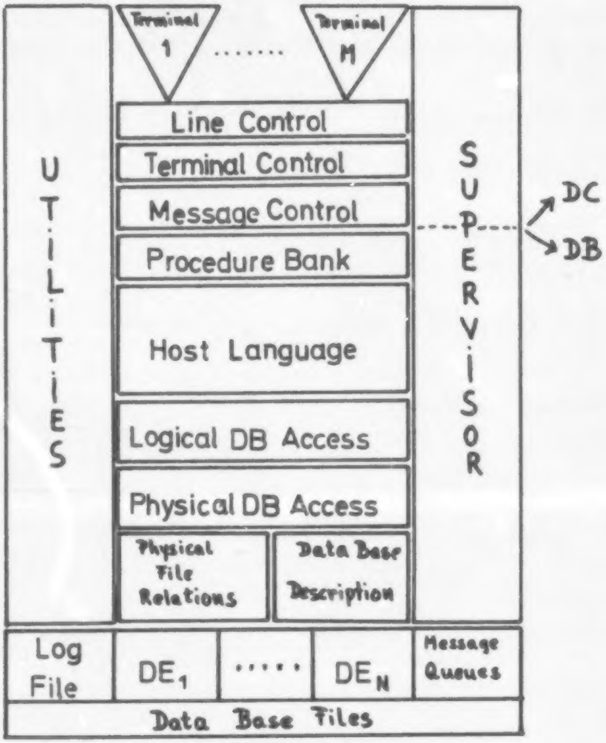

Terminal 1
Terminal M
Line Control
Terminal Control
Message Control
Procedure Bank
Host Language
Logical DB Access
Physical DB Access
Physical File Relations
Data Base Description
UTILITIES
SUPERVISOR
DC
DB
Log File
DE$_1$
DE$_N$
Message Queues
Data Base Files

Fig. 7.2: DB/DC-system components

adequate to our problem are more complexe operations as *select, project* or *join*. *Select* defines always a subset of a relation by indicating through a selective condition which tupels should be selected and which not. So

S;ANR=166

defines the following subset from S:

210	166	a
294	166	b

This subset shows all students advised by advisor number 166. As *select* takes tupels from a relation, *project* asks for domains. So if we write for projection (% as symbol)

S%ANR

we achieve a domain set where duplicates are suppressed as follows:

166
169
181

This set is an answer to the question which advisors are advising students.

Join finally combines two relations in a restricted way as it is described by a synchronizational condition which tupels from all combinations should be taken (symbol is x). So

AxS;ANR:A=ANR=S

delivers the following new and enhanced relation:

166	A	4	1	210	166	a
166	A	4	1	294	166	b
169	D	1	2	280	169	c
181	C	5	1	298	181	c

This is a list expressed of all advisors really advising, showing data for their students. Another projection would delete not needed domains.

The query: *Find students advised by advisors in faculty 2 by SNR, name and ANR* can now be formulated by relational operators.

The following sequence of relational operators defines the relation, answering this question, where interim relations occur. Their domains are only named by indices.

X = ADVISOR; FAC=2
X1 = X%ANR,FAC
X2 = X1xSTUDENT;ANR:S=ANR:A
X3 = X2%3,4,5
LIST: X3

This sequence of algebraic expressions is no longer of algorithmic nature and a quite different approach to data base evaluation than usual. Obviously all auxiliary relations can be eliminated by insertion and we achieve one single algebraic representing the structure of our query:

Query:

Find all students advised by advisor
in faculty 2 (Fac=2) by SNR, Name, ANR

Data Base:

$\Gamma_1 : V_1 = \text{Advisor}(V_1\{ANR\}, V_1\{ANAME\}, V_1\{Room\}, V_1\{FAC\})$

$\Gamma_2 : V_2 = \text{Student}(V_2\{SNR\}, V_2\{ANR\}, V_2\{Name\})$

Tupel Variables:

$V_1 \in$ Advisor $V_2 \in$ Student

Query:

⋮

$\{V_2[SNR], V_2[ANR], V_2[Name];$

$\Gamma_2 \wedge \exists \Gamma_1 \forall \Gamma_2 \; (V_2\{ANR\} = V_1\{ANR\})\}$

wherein $\Gamma_2 \in$ Student

$\Gamma_1 \in$ Advisor $\wedge \; V_1\{Fac\} = 2$

Fig. 8: Predicate calculus data base use

LIST: ((ADVISOR;FAC=2)%ANR,FAC)x(STUDENT;ANR:S=ANR:Y)
%(SNR,ANR,NAME)

Such an algebra of considering data base files as algebraic entities opens a lot of algebraic questions such as completeness, usefulness of definitions and operations used, etc. If the algebraic approach combines sets given as entities and defined by existence through appropriate operations the predicate calculus defines predicates by which set are defined being responds to a query. Such a predicate calculus query as showed in Figure 8 is a combination of a logical notation of elements (symbols for AND, OR, NOT, quantifiers, variables, relation names and domain names) as used in formal logic. Also there occur research problems especially with respect to formal and natural language problems if the predicate calculus is considered as the unique skeleton of a natural language.

Literature

(1) Date, C.J.: An Introduction to DB System. Addison-Wesley 1975.
(2) Berztiss, A.T.: Data Structures. Academic Press, 4 edition 1973.
(3) Rustin, R. (Eds.): Data Base Systems. Prentice Hall 1972.
(4) Cagan, C.: Data Management System. Melville 1973.
(5) Codd, E.F.: A Relational Model for Data for Large Shared Data Bases. CACM 13/6 (1970).
(6) Lutz, Th.: Die Datenbank im Informationssystem. Oldenburg 1971.
(7) Lutz, Th.: Das computerorientierte Informationssystem. de Gruyter 1973.
(8) Senko, M.E.: Data Structure and Accessing in Data Base Systems. IBM Systems Journal Vol. 12, Nr. 1, 1973.
(9) Lutz, Th.: Die Mathematik der Datenbank. IBM Nachrichten 225 ff., 1975.
(10) Lutz, Th.: Informationssysteme und Datenbanken. IBM Nachrichten 216 bis 218, 1973/74.
(11) Codasyl Programming Language Committee. Data Base Task Group Report. April 1971.

Handbooks and Manuals for IMS and GIS where used.

Data Documentation – an Overview

Christoph Schäfer, Frankfurt/Main

1. Definitions

Data documentation provides *direct* answers to questions "*how* is something". *Literature documentation* (or bibliographic documentation) provides only *indirect* answers "*where* to find information about".

a) Data
According to a survey (1) 'Data and the Chemist' by ASLIB, U.K. approx. 50% of interviewed chemists considered data as a synonym word for *facts*, 30% for 'primarily numerical' information and 20% for a fairly meaningless or pompous term (nonsense word).

Here the first definition is supported, interpreting the term facts in a broad sense, i.e. facts being measured or observed, quantitative or qualitative, numerical or non-numerical, scientific or non-scientific.

b) Documentation
Touloukian (2) complained about the confusion in terminology. He proposed to refer to those centres which store, retrieve, and disseminate raw numerical data as *information centres*, in contrast to *documentation centres* which process bibliographic or documentary facts. "The usage of the term *information* in reference to bibliographic and documentary facts should be avoided".

These proposals seem to be too contrary to the historically developed terminology in order to be internationally agreed upon. Data documentation can be defined as compilation, retrieval, and dissemination of numerical and/or non-numerical facts other than bibliographical data.

2. Data categorization

Data can be categorized in various ways; the following scheme is proposed, being a rather pragmatic approach:

A. Non-numerical data

Correlation of two or more factual units.
Examples: directory of addresses; list of plants or animals and their geographical distribution.

B. Numerical data

a) Physical-chemical properties of substances

These data are reproducible and independent of time and location.

Examples: Melting points of metals, surface tension of liquids etc.

b) *Industrial* data, *engineering* data or *material* data

These data depend on the present and past of the matter or on the geometry of the pieces.

Examples: Reflexion-rate of light on a metal surface.

c) Property data of natural products or phenomena, as in geosciences, astrophysics and life sciences.

Most of these data are time and location dependent and therefore not reproducible. Certain data are time independent but location dependent e.g. in geology.

Examples: Temperature of air in the atmosphere, solar flares, blood pressure of mammals.

d) Social and economic data as in sociology, education, politics, economics, finances, administration.

Examples: Age of persons, income of professions, number of students.

3. Data treatment or handling

a) Generation of data

The scientists are responsible for the generation of data, they have to decide on the type of data to be created and on the method and procedure to be applied. Data treatment at this stage is connected with a variety of procedures.

Examples:
- analysis and choice of experimental, observational and statistical methods.
- choice and calibration of instruments.
- systematic error control.
- analysis of the applicability of models.
- analysis of the procedure of data calculation.
- method of reduction, interpolation, extrapolation, and correlation of data.

The final result of data generation is in general the preparation of printed documents in the form of books, reports, or journal articles with numerical or non-numerical tables, charts, graphs, curves.

b) Compilation of data

The end products of data generation (= documents with facts) are the starting points of data documentation. The extraction, compilation and repackaging

96

of raw facts from the above mentioned printed documents can be considered as part of data documentation.

c) Evaluation of data

The critical evaluation of any information, textual or factual, should not be a task of documentalists and information specialists (except in their own field of competence: documentation of information processes). Evaluation can be exercised only by high qualified experts in very restricted fields of knowledge. Critically evaluated information (data) is compressed *and* refined information, i.e. information of high density and high quality.

d) Dissemination of data

As well as the compilation of data the dissemination falls within the scope of documentation/information.

The compilation and evaluation of data and their dissemination to users are the central task of information centres.

Libraries should facilitate the accessibility to raw and evaluated data, compilations, and to directories of data sources by considering data handbooks and compilations in their budgets adequately.

Summary of information activities

Activity	Product	Implemented by
Information generation ↓	→ document	research scientists
Collection ↓	→ bibliography	documentalist
Indexing, abstracting ↓	→ annotated bibliography, abstract	documentalist
Extraction ↓	concepts → survey state-of-the-art facts → uncritical data compilation	documentalist*
Evaluation	→ critical review critical data compilation correlation recommendation	research scientist

---

* data extraction and repackaging has been implemented by research scientists so far. In future this could and should be the task of scientific information experts. It is anticipated that the preparation of abstracts by scientific information experts will become more and more superfluous, due to the increasing number and improved quality of author abstracts and of the dominating role of the 'abstracting factories', the big abstracting and indexing services.

It would be useful if scientific documentalists would substitute the preparation of abstracts by extracting raw, unevaluated data from the literature for the preparation of data compilations.

4. Data documentation

Above four main activities were distinguished within the area of data documentation:

a) Extraction, compilation and repackaging of published or otherwise recorded data
b) Improvement of accessibility to data compilations and data sources
c) Storage and retrieval of data
d) Dissemination of data

The following problem areas need further exploration in order to improve the efficiency of all aspects of data documentation:

a) Systems interconnection between bibliographical information systems and data compilation, evaluation, and retrieval
b) Standardization of data recording and of data presentation in handbooks, tables etc.
c) Terminology and classification of data
d) Proper data recording immediately after data generation
e) Computer storage of data
f) Accessibility and dissemination of data
g) Identification of 'hidden' data compilations
h) Data needs of potential users

Literature

(1) Slater, M.; A. Osborn; A. Presanis: Data and the Chemist, Aslib Occasional Publication No. 10, ASLIB, London, 1972, 82 pp.
(2) Touloukian, Y.S.: Reference Data in Thermophysics. In: CODATA Bulletin No. 14 (1974), CODATA, Paris, p. 138–152.

Confidentiality and Security in Statistical Data Banks

Jan Schlörer, Ulm

The issue of computer privacy and security in connection with personal records – records containing full identification data of the persons concerned – is frequently discussed. It is less commonly known that anonymous statistical data on persons, too, may raise problems of confidentiality and security. As with personal records, the increasing use of computers as a tool for storing and evaluating statistical data has aggravated and, in part, even generated the risks to be discussed.

1. Definitions and terms

Before going into detail, some definitions and terms will be helpful. First of all, we need some classification of record and data types suitable for our theme. There are:

(a) *Personal* records, which contain data on exactly one person, say, on John Doe, including identification data such as name, address, or – in times to come – some universal identifier. There is a 1:1-relation between records and persons.

(b) So-called *anonymous* records. Each record contains data on exactly one person; the 1:1-relation between records and persons is preserved. But the conventional identifiers – name etc. – are missing now.

(c) *Aggregate* or aggregate statistical data. The best known example are statistical tables. The 1:1-relation between records and persons exists no longer. Aggregate data, of course, are anonymous.

Data on organizations may be rubricated along the same lines. They are subject to analogous confidentiality risks, which will not be explicitly mentioned in what follows.

To prevent misunderstandings some basic statistical terms and symbols will be recapitulated. Some examples for the meaning of *"variable"* and *"variate"* are:

Variable	Type of variable	Variates
Sex	Qualitative	Male, female
Number of children	Discrete quantitative	0, 1, 2 children etc.
Age of life	Continuous quantitative	11.17, 53.875 years etc.

Continuous quantitative variables may be classified, resulting in "classes" (say, 10 – < 20, 20 – < 30 years etc.). Sometimes with other types of variables, too, the term "class" (instead of "variate") is employed. The expressions "property" or "attribute" (of some Mr. X, whose record we want to identify) also mean "variate".

The following symbols will be used:

N = Size of a population (resp. of a record collection, which is a population; note that, strictly speaking, a population or sample does not consist of persons or whatever, but of observed (or observable) variates only).

n = Size of a sample (resp. of a record collection which is a sample).

f_i = Absolute frequency of variate (class) i. Often we will need slightly different symbols, but the letter f always signifies, that an absolute frequency is involved.

p_i = Relative frequency of variate (class) $i = f_i/N$ or f_i/n.

A *statistical data bank* will be assumed to consist of a collection of logical (not necessarily physical) records, containing one record per person. The logical record may be split and distributed over several files with differing internal structure (e.g. (34, 35)), but the physical data structure will be neglected. All logical records are assumed to be equally formatted and to contain variables with disjunct classes only (only one variate per variable and record), because for statistical purposes this is the most important situation. An information system of a statistical data bank is a set of one or more data banks, plus the respective system of evaluation programs. Strictly speaking, such a system contains many more elements (see, e.g. (37)), but they are of minor importance for our considerations. The user is an element of any information system too.

A well designed query language for statistical data banks should offer, among others, the *logical operations* AND ($\wedge$), OR ($\vee$), and NOT ($\neg$), as well as a facility to generate and store new logical operands (see below), which is equivalent to an option to construct new composite variables (e.g. (34, 35)). The user then may enter some logical condition (an inquiry or query; input) into the system, which will return as output the number of records fulfilling the condition entered. This facility is not only important for the bona fide user, but at the same time constitutes a central tool for dialogue "snooping techniques".

We will denote logical *conditions* by capital letters, which also symbolize the set of records defined by the equally named logical condition. Individual logical *operands* will be denoted by lower case letters, since they mostly represent variates. The symbols $\wedge$, $\vee$, and $\neg$ are called logical *operators*. The number of records fulfilling some logical condition, say: A, will be denoted by f(A). f(A) is an absolute frequency, and at the same time denotes the cardinality of the subset A of the records contained in the statistical data bank. If, for example, the inquiry

$$A \quad = a_1 \wedge a_2$$

is entered, where a_1 = male, and a_2 = female, then, of course,

$$f(A) = f(a_1 \wedge a_2) = 0,$$

since no record can contain both entries, male and female. So the set A is empty.

The facility of generating and storing new logical operands simply means, that the user may enter some logical condition (say: $a_3 \wedge a_4$), store the result under some name (say, C), and then may use C instead of $a_3 \wedge a_4$ in formulating a later query – say, $C \wedge (\neg a_5)$, instead of $(a_3 \wedge a_4) \vee (\neg a_5)$.

2. Confidentiality risks

2.1 Additional knowledge

Of the three data types mentioned (personal records, anonymous records, aggregate data) we are only interested in the last two ones. Regarding confidentiality, there is a principal difference between personal records on the one hand and anonymous records plus aggregate data on the other hand.

If you can acquire a personal record of Mr. X and can understand it (coding (24), cryptography and the like are disregarded in this context), then you can use the information contained in the record immediately and come to some decision concerning Mr. X based on that information.

With anonymous records or aggregate data this is not feasible. They first have to be de-anonymized. To this end one always needs (16, 18, 20, 21, 24–26, 30)

(a) some additional knowledge (mostly on Mr. X, but in special situations on other persons, too), which

(b) may be matched in some way to information contained in the statistical data bank.

In the most typical case you undertake some kind of record linkage between the anonymous records, which you try to identify, and the imaginary record containing your additional knowledge on Mr. X, to find out which of the anonymous records matches your additional knowledge, and therefore belongs to Mr. X (16, 18, 20, 21, 24). There are more complicated situations (16, 20, 25, 26, 30), but the requirement of additional knowledge, which corresponds in some way to the information stored in the statistical data bank, is generally valid.

The requirement of matching may be put the other way round. If anonymous records or aggregate data contain only information which is available nowhere else, then there is no confidentiality risk whatsoever. But practically each and every statistical record on human beings contains some so-called administrative, demographic or social status data: sex, personal status, age of life, and so on, which are generally known or may be easily obtained from a variety of sources.

2.2 Aggregate data

Another appropriate header to this section would be "off line dialogue". An off line dialogue with a statistical data bank is given, if a customer asks a statistical office for a set of tables which is then generated according to the customer's specifications in the office. In on line dialogue the user may enter his inquiries directly into the system.

Both, on line dialogue and off line dialogue produce usually aggregate data as output. But in on line dialogue the questions may be put such that – if indirectly – complete anonymous records, instead of aggregate data only, may be inspected by the user. Off line dialogue scarely will provide a user with similar opportunities. He usually will gain, at best, some fragmentary information on identifiable respondents. Moreover, the additional knowledge needed tends to be – but is not necessarily – relatively complex. This does not imply that there are no risks in off line dialogue, as Fellegi, Nargundkar, Saveland, and others at Statistics Canada clearly have demonstrated (16, 17, 25, 26):

(a) It may be possible to deduce properties of identifiable respondents solely from a statistical table or from a combination of such tables.

(b) Theoretically one might monitor all tables being published, say, by a statistical office, to ensure that no table or combination of tables issued so far, results in disclosure on identifiable respondents (16). But in practice this would imply an insurmountable amount of work, and, worse, it would reduce, in the long run, the options for generating new tables, as they are requested by the office's customers – more and more tables would become "forbidden".

(c) The main risk resides in the true absolute frequencies, which either appear in tables, or may be deduced from derived quantities (percentages, arithmetic means, etc.).

These considerations together with the quite tough-worded Canadian Statistics Act from 1970 have led Statistics Canada to stop publication of any true absolute frequencies, beginning with the 1971 Census of Canada. The technique adopted is called random-rounding, and will be discussed later on.

All known snooping techniques (at least the more important ones, which – ideally – are "deterministic" (32)) rely to some extent on the formation of subsets of records, whose cardiualities (absolute frequencies) an intruder must learn in order to succeed. This applies to both off and on line dialogue; in fact, the general scheme developed by Fellegi (16), Nargundkar, and Saveland (25, 26) will, after slight modification, accomodate the on line techniques as well (32), even if it does not tell anything about the algorithms needed. So the critical importance of absolute frequencies applies to on line dialogue, too.

2.3 Anonymous records: underlying assumptions

Anonymous records and on line dialogue are considerably more vulnerable than aggregate data and off line dialogue. If you get hold of some anonymous records and manage to identify the one belonging to Mr. X, then you have transformed it into a personal record. But physical capture of the record is not necessary. Even if the query algorithm of some statictical data bank information system denies any access to complete anonymous records, the normal user sometimes may identify and – step by step – retrieve such records by on line dialogue. To simplify things we will make a number of assumptions; some of them will be dropped later:

(a) The statistical data bank contains a population of anonymous records (implying that the intruder knows that a record on Mr. X must be stored in it), it contains error-free data, and it contains enough variables fit for use as additional knowledge to permit identifications.

(b) The intruder behaves like any normal user (no stealing of tapes and the like); his additional knowledge is error-free.

(c) The query algorithm offers AND, OR, and NOT; it permits to generate and store new logical operands, and it returns true absolute frequencies to the user.

2.4 Anonymous records: identification

A central question is: how much additional knowledge on Mr. X do we need in order to identify his record? Unfortunately this is a question which may not be answered once and for all. If the variables, whose variates are contained in anonymous records and may serve as additional knowledge, were statistically independent, then a general rule easily could be given (21). But there are not only dependences; moreover, different sets of records contain different sets of variables, the correlations and contingencies within the same set of variables depend on the population we are studying; finally, within the same population, the dependences are time-dependent. So we may solve the question only empirically for a given set of anonymous records.

Further, one never may predict exactly what additional knowledge, and which external information sources an intruder might have at his disposal. The judgment, whether some variable is "dangerous" (suitable for additional knowledge) or "innocent", will only be a more or less well informed guess, for which the most important criteria seem to be the ease of access to some piece of information, and its time-invariance – your date of birth, e.g., remains constant for life, while your place of residence may change frequently.

So empirical data drawn from one statistical data bank are not fully transferable to a different set of statistical data. I have done an identification experiment, using a statistical data bank containing 31 465 anonymous records stemming from a pilot study in preventive health care in Baden Württemberg (11). The experiment involved simulated identifications only, and no actual breach of anonymity occurred. The details have been described elsewhere (30), so we will only look at some of the results.

It turned out that there was a high variability in the amount of additional knowledge needed. Taking the relative frequencies p_i of the properties employed as additional knowledge as a yardstick, with some records very much, with other records relatively little additional knowledge was necessary. More than 10% of the records required less additional knowledge for their identification than would have been expected assuming statistically independent variables.

A certain amount of additional knowledge would do for identification in a certain percentage of the cases. On the basis of these data it therefore was possible to calculate the number of attributes (each attribute being a variate of a different variable), which would do for identification in a certain percentage of trials, given the size N of the population in question, and assuming that *all* attributes had the *same* relative frequency p_X. To achieve identification in about 85% of the cases, the following numbers of properties would have been needed (the numbers are rounded up to the next higher integer throughout):

Px N	0.7	0.5	0.3	0.2	0.1	0.05
31 465	42	22	13	10	7	5
1 000 000	52	27	16	12	8	7
60 000 000	64	33	19	14	10	8

To succeed in only 5% of the trials, less additional knowledge is needed, of course:

Px N	0.7	0.5	0.3	0.2	0.1	0.05
31 465	28	15	9	7	5	4
1 000 000	38	20	11	9	6	5
60 000 000	49	25	15	11	8	6

These are not very high figures; they increase not with N but only with log N. So identifying Mr. X among 60 million people does not require that much more information than finding him out among, say, only 100 000 people.

It must be repeated that figures of this kind are not fully transferable to other statistical data banks. They should be empirically determined for each set of anonymous records where it is important to know them more closely. On the other hand, the experiment probably gives a valid impression of at least the order of magnitude of the additional knowledge needed to identify people, or better, their records within a given population.

2.5 Anonymous records: retrieval

By now it is clear that – in the hands of an intruder, who has captured the physical data, and has sufficient additional knowledge – a set of anonymous records, which contain enough information matching the additional knowledge, is not very much different from a set of personal records. Without an opportunity to take the data away, but depending instead on on line dialogue, an intruder faces a more tedious, but not principally different situation, as long as the assumptions made in section 2.3 hold.

At this point we must mention the possibility of output restrictions for small and large absolute frequencies:

Z_{min}	Minimum absolute frequency, which a system returns to the user. An inquiry Y leading to an $f(Y) < Z_{min}$ will be canceled; any logical operands resulting from this inquiry will be erased.
Z_{max}	$N - Z_{min}$, where N = number of anonymous records in the statistical data bank (which is still assumed to be a population!). Largest absolute frequency which the user is allowed to see. Functions analogous to Z_{min}; for the necessity of Z_{max} see (30).

All on line retrieval techniques known so far (20, 25, 27, 29–31) rely on absolute frequencies for identification of the anonymous records in question. In other respects they differ, three of which we will mention (a more detailed review appears in (32)):

(1) For retrieval they may rely on
  (a) derived statistical measures, or on
  (b) absolute frequencies, as for identification

(2) For identification they may rely on subsets of records of
  (a) cardinality > 1 (no full identification)
  (b) cardinality = 1 (full identification)

(3) To output restrictions (Z_{min}, Z_{max}) they may be
  (a) sensitive
  (b) immune

It is important to realize that (2b) and (3b) are not mutually exclusive – the set of cardinality 1 must not necessarily appear as output, but may be derived by arithmetic operations (e.g., subtraction) from other output figures (a concept called "residual disclosure" by Fellegi (16) and "disclosure among cells" by Nargundkar and Saveland (25, 26)). The most dangerous techniques, of course, are those employing full identification (2b), which at the same time are immune to output restrictions (3b):

(1) Palme's technique (27) which employs arithmetic means ($\overline{x}$) or sums of variates (Σx_i) for retrieval. It will retrieve quantitative variables only. The

quantile technique sketched by Nargundkar and Saveland ((25), p. 49) is applicable to quantitative and rank order variables, but not fully immune to output restrictions.

(2) The tracker technique (29–31), which employs absolute frequencies not only for identification, but also for retrieval. It is applicable to all types of variables; it only bocomes tedious to impossible, where it encounters unclassified continuous quantitative variables, which the user is not allowed to classify (which rarely will occur).

As an illustration examples for two techniques will follow: (a) The first of two algorithms given by Hoffman and Miller (20) – this was, to my knowledge, the first detailed description of an on line retrieval technique; (b) one variation of the tracker technique. Let the following 5 attributes identify our Mr. X:

Variable	Variate	Code
Residence	Rothenburg	01
Age	20 – < 25 years	02
Sex	male	03
Number of children	2	04
Line of employment	Housing & health care	05

The variables, whose variates we want to know for Mr. X, are

Has been convicted of a felony	yes	06
	no	07
Has a record of psychical illness	yes	08
	no	09

The algorithm of Hoffman and Miller then goes:

Logical condition Y (input)	f(Y) (output)	Code of newly generated and stored operand, if any
$01\wedge02\wedge03\wedge04\wedge05$	1	10
$10\wedge06$	0	–
$10\wedge07$	1	–
$10\wedge08$	1	–
$10\wedge09$	0	–

We obviously may conclude that Mr. X has not been convicted of a felony, but has had some psychical illness. Using one of the variations of the tracker technique, the following picture emerges:

04	7173	– (A_1, see sect. 3.4)
$(\neg(01\wedge 02\wedge 03\wedge 05))\wedge 04$	7172	10 (tracker)
$(10\vee 06)\wedge 04$	7172	–
$(10\vee 07)\wedge 04$	7173	–
$(10\vee 08)\wedge 04$	7173	–
$(10\vee 09)\wedge 04$	7172	–

Here the interpretation is less obvious. The subset of cardinality 1 shows up as the *difference* between the f(Y) of the first two inquiries, the first of which contains Mr. X's record, whereas the second one does not do so – and that's the only difference between the subsets of records defined by both queries (for details see (30)). The same goes for the later inquiries; whereever f(Y) = f(04) = 7173, Mr. X's record is contained in the respective subset and fulfils the logical condition Y. So our conclusions are the same as before.

2.6 Anonymous records: dropping the underlying assumptions

Of the assumptions made in section 2.3 we now will drop those concerning the statistical data bank (a) and the intruder (b); the query algorithm will be dealt with later on (sections 3.3/3.4). More details may be found in (16, 18, 24, 29, 32).

(a) An anonymous record containing too little information suitable as additional knowledge will permit no identifications.

Data banks containing only samples (size = n, $n < N$) may still permit identifications. But this usually requires more additional knowledge than identification form the respective population (29), except if (which rarely will be the case) the intruder knows a priori that Mr. X has no *doppelgänger* within the population (16) – otherwise the probability that some record does belong to Mr. X, and not to some *doppelgänger* may only be sufficiently increased by employing surplus additional knowledge. Moreover, the prospects of spotting a Mr. X within a sample of size n are only $\leqq n/N$. The risks just described are practically meaningless for samples with a low ratio n/N, whose composition is kept confidential (16, 18, 24, 29).

Errors in the data, too, tend to reduce the intruder's chances (18, 24). But again, with surplus additional knowledge, the errors may be circumvented and identifications achieved (29). The information then retrieved also is unlikely to be errorfree, but – as Nargundkar and Saveland have pointed out (25) – this is not a guarantee against misuse. With a high percentage of errors, about which the intruder knows, the risk certainly will become low, but the statistical value of the data will do also.

(b) Errors in the intruder's additional knowledge have the same consequences as errors in the data. But you normally have no influence upon the intruder's additional knowledge, so you better assume that his information is correct.

If the intruder captures the physical data (and has got a computer of his own), his task may become easier, but the situation is not principally different from the one described in section 2.4 above.

3. Countermeasures

3.1 Statistical versus dossier data banks

"Dossier" here is employed without a derogatory notion, "dossier data bank" is meant to be shorthand for a data bank containing personal records. Dossier and statistical data banks overlap to a considerable extent, but there is an important functional distinction between them (36, 37): Dossier data banks serve to prepare decisions concerning the identifiable persons whose records are stored. Accuracy of the individual records and of the output from a record inspected are essential. Statistical data banks serve to gain information on groups and phenomena of large numbers. As long as the general conclusions to be drawn are not too heavily impaired, the contents of the individual records and the lower digits of the output are not that important; identification of individual records, if at all, is necessary for reasons of data organization only (e.g., for longitudinal studies 1)). This implies that – in addition to "conventional" countermeasures, which are applicable to both dossier and statistical data banks (and which, strictly speaking, mostly are not yet that conventional) – there are safeguards which are specific to statistical data banks. A rough threat-countermeasure matrix for statistical data banks appears in table 1.

Statistical data banks may well require "conventional" countermeasures – e.g., cryptography in order to protect the physical anonymous records. It also may be necessary to control the access to a statistical data bank, but at least for statistical data banks of interest to the public this is a cure, which I personally deem not much better than the disease. For reviews of these and many other "conventional" safeguards see, e.g., (13, 19, 36).

For statistical as well as for dossier data banks one should not try to rely on single, if seemingly very effective safeguards; always a spectrum of partially overlapping techniques should be employed. Moreover, the statement of Han-

---

1) For longitudinal statistical data the first problem to be solved is to transform the personal records into "normal" anonymous records. To this end a class of ingenious split-and-link techniques have been developed, details on which are given in (5, 8, 24).

sen (18), that "if *any* statistics are to be published nondisclosure cannot be absolute" has come here to stay.

Threat Group of counter-measures	Computer crime		"Legal" attempts to capture the data (sub poena; modification of laws; political pressure etc.)
	Techniques relying on off line or on line dialogue (the intruder behaves like any normal user)	Other forms (capture of physical files by external intruders or systems personnel etc.)	
Data transformations (section 3.2)	+	+	+
Modification of the query algorithm or output (sections 3.3/3.4)	+	φ	φ
"Conventional" countermeasures (applicable to dossier as well as to statistical data banks)	φ	+	φ

Table 1:
Main groups of threats and countermeasures for statistical data banks. + signifies that the respective safeguards may exert some effect, φ that they do not.

3.2 Data transformations

Data tranformations may, to some degree, protect statistical data against all kinds of threats. The data are transformed such that identifications become difficult or impossible, whereas statistical evaluation remains feasible. The drawback is that any such transformation means loss of information (28), and therefore reduces the options for evaluating the data. So data tranformations – excluding, perhaps, the use of samples instead of populations – have only limited applications and are no panacea. Nevertheless it is worth while to know their outlines; there are situations where data may be modified just slightly, resulting not in full but in enhanced security.

(a) Resorting to samples of size $n \ll N$, instead of storing populations; the transformation consists in eliminating whole records. The effects thereof have already been mentioned in section 2.6.

(b) Elimination of "dangerous" variables. The reduction of the number of variables (fields) of a given record type, which could serve as additional knowledge, will reduce the risk of identifications correspondingly. The purposes, for which the data were collected, often simply will not allow this escape. It seems, however, in order to screen the obviously "dangerous" variables with an extremely jaundiced eye, to make sure that they really need to be collected and stored.

(c) Reducing the number of variates, that is, coarsening the classifications of "dangerous" variables (28), will increase the p_i of the remaining variates and thereby decrease their discriminating or identifying power, as may be seen from the figures in section 2.4. Again, this will not always be feasible. But it is scarcely always necessary to know, e.g., a respondent's date of birth; the age in years, sometimes even the age group in 2- or 5-year classes, will do perfectly. This especially holds for statistical mass data on populations, where mainly descriptive evaluation is intended.

(d) Controlled introduction of random errors into the data is not too secure a device, as long as the error quote remains low (see section 2.6). Moreover, the options for two-dimensional (not to speak of multivariate) analysis rapidly deteriorate with increasing error rate (4, 18); to date there seem to be no proved methods for two-dimensional evaluation of error-inoculated data. The effect of low error rates, if combined with different safeguards, remains to be studied. Meanwhile a subset of these techniques (9, 10, 23, 38) already has an important practical application: they may serve to estimate absolute and relative frequencies of variates, which may not be found out by more direct ways (e.g., percentages of abortions and the like); Here the errors are introduced at interview time; already the interviewer will never come to know anything about the true situation of his respondents.

(e) Micro- or partial aggregation (14, 15). When suitable aggregation criteria are chosen, two- and higher-dimensional (especially regression) analysis remain possible, though with somewhat reduced efficiency. The suitability of a set of aggregation criteria, however, depends on the questions to be asked. Different users of the same data might wish differently aggregated data, which – if granted – would tend to reduce the protection effect.

The five transformation types mentioned so far are relatively well understood. I should like to sketch one further possibility, two-dimensional transformation, even though its practical value cannot yet be determined, since to date only its outlines are known. The idea is to preserve the two-dimensional relations within a transformed set of records (all equally formatted), while the original records are lost. The most simple structure allowing a transformation of this kind is illustrated in the example below:

Record Nr.	Sex	Marital status	Mentally ill?		Sex	Marital status	Mentally ill?
1	female	single	no		female	single	yes
2	female	married	yes	⟶	female	married	no
3	male	single	yes	⟵	male	single	no
4	male	married	no		male	married	yes

Note that any (horizontal) pairs of variates on both sides of the diagram are preserved: there is one record with the entries "married/mentally ill" on either side, and so on. But the records in toto on both sides are different. The general structure underlying this example is:

Record Nr.	Obligate	Optional
1 2 3 4	$\begin{bmatrix} a & a & a \\ a & b & b \\ b & a & b \\ b & b & a \end{bmatrix}^1$	$\begin{pmatrix} a \\ a \\ b \\ b \end{pmatrix}^\infty \begin{pmatrix} a \\ b \\ a \\ b \end{pmatrix}^\infty \begin{pmatrix} a \\ b \\ b \\ a \end{pmatrix}^\infty \begin{pmatrix} a \\ a \\ a \\ a \end{pmatrix}^\infty$

Each vertical column represents one variable, while a and b denote two different variates – it is unimportant which ones – of the respective variable. You may determine yourself that you just have to perform, e.g., the transformation

$$(a\ b\ b\ a)^T \rightarrow (b\ a\ a\ b)^T$$

whereever a column of the structure $(a\ b\ b\ a)^T$ occurs, in order to preserve any two-dimensionalc contingencies – but the original records have gone after that. There are blocks of this kind comprising more than 4 (e.g., 6, 8 and so on) records, but they become rarer with the increase of the number of records involved. The principle then would be, to scan a given set of records for blocks of records which may be transformed and, as soon as one is encountered, to transform it. Suitable blocks will become more frequent with increasing number of records per data bank, they will become rarer with increasing number of variables per record type and with increasing number of variates per variable. The exact probability for the occurrence of these blocks is still only known for very special cases; even if it would prove sufficiently high for practical purposes, there would remain the formidable problem of finding a suitable transformation algorithm. There are other restrictions: Three- and higher-dimensional relations are lost; continuous type variables probably had to be relatively roughly classified before the transformation – thereafter one only might further coarsen the classifications, but not refine them again.

3.3 Modification of query algorithm; threat monitoring

While output modification concern both off and on line dialogue with a statistical data bank, modifications of the query algorithm are interesting mainly for on line dialogue.

As long as the bona fide user of a statistical data bank is, in on line dialogue, to enjoy the facility to generate new, composite variables (to this end he must be permitted to generate and store new logical operands), restriction of the query language's syntax is no practical means of protection. The building of trackers as well as of the identifying conditions for Palme's technique may only be prevented by synteax restrictions which are so severe, that the facility to generate new variables is lost (31).

There remains another possibility. It is, if only via a crude worst case analysis, possible to estimate the identifying power of the inquiries entered into the system (33). Of the on line retrieval techniques known so far, which are capable of retrieving complete records (the other techniques are more susceptible to output modifications), none avoids identifying conditions

$$a_1 * a_2 * \ldots * a_k$$

where $*$ denotes any logical operators, while the a_i $(i = 1, 2 \ldots k)$ are variates of k different variables, which, if entered in the form

$$A = a_1 \wedge a_2 \wedge \ldots \wedge a_k$$

would result in $f(A) = 1$ 2). Monitoring the use of the "dangerous" variables employed by the logical conditions entered into the system, and preventing the use of more than one classification per variable and inquiry, may largely cut off the formulation of identifying conditions. This by itself will not do; in addition output modifications, if probably only slight ones, will be needed to render possible uncontrolled on line access to statictical data banks, which contain anonymous records open to identification.

---

2) Meanwhile is has turned out that there are – if cumbersome – ways to circumvent identifying conditions. One therefore, for "dangerous" variables, whose classifications the user wishes to change, may only offer standard classifications, which contain no overlapping, but only "hierarchically" splitted resp. merged classes (33).

3.4 Output modification (rounding)

The modification of output concerns, in first line, true absolute frequencies. But the "back-doors" – derived quantities, from which absolute frequencies may be deduced (e.g., percentages and many others (17, 25)) – must also be watched. The method of choice is rounding. Instead of the true frequency f a rounded value

$$\hat{f} = g(f)$$

is returned to the user, where $\hat{f}$ usually (but not necessarily) will be some multiple of the rounding base b, and g is some rounding function. Any such function should meet the requirement, that it may be published without serious loss of protection effect: firstly, a safeguard, which itself needs to be kept secret, is rather vulnerable (6); secondly, the user must be able to estimate the information loss, by which he has to pay for his free access to a statistical data bank. So 1:1-functions are out of question; we are left with:

(a) Random-rounding (1: many – "functions"). For each value f a pseudo-random number will decide ad hoc, whether $\hat{f}$ will be the next higher or the next lower multiple of the rounding base b. The technique has been developed and implemented, using a rounding base $b = 5$, by Statistics Canada (17, 25, 26).

(b) Determinate rounding (many: 1 – functions). The rounding rule g predetermines the $\hat{f}$ for each f.

Both methods have in common that protection effect and loss of information go up with the rounding base b. There is also a problem of deroundability of tables for both techniques. A deroundable table is a table consisting of $\hat{f}$ values only, from which one may deduce all or part of the true f values, if one knows the rounding rule g. For systematic rounding (a subspecies of determinate rounding, see below) with $b = 5$ Fellegi and Phillips (17) have given an example for a deroundable table. For random-rounding with $b = 5$ the following table is a case in point:

Random-rounded table ($b = 5$) ((containing $\hat{f}$ values only))

			Σ
	10	10	25
	10	10	30
Σ	25	30	60

Values, which according to random-rounding algorithm (14, 19, 20) may underly the rounded values:

			Σ
	6–14	6–14	21–29
	6–14	6–14	26–34
Σ	21–29	26–34	56–64

True or derounded f values (the only table, which is possible, given the values in the middle):

			Σ
	14	14	28
	14	14	28
Σ	28	28	56

The relative frequency of deroundable (among all possible) tables depends on the size (degrees of freedom) of the table and on the rounding base. Both being equal, deroundable tables probably are rarer for random-rounding than for systematic rounding (at Statistics Canada originally it was hoped that random-rounding could avoid them completely (17)), but their exact rate of occurrence is not yet known. If one is ready to accept a somewhat increased rounding error, one may circumvent the problem by calculating marginal sums as sums of $\hat{f}$ values (instead of rounding the marginal f values directly).

Random-rounding is preferable to determinate rounding in its ability to achieve zero bias (25, 26). The distribution of last digits of the occurring true absolute frequencies is unpredictable – in fact there are examples that they are not always equidistributed (25). So zero bias can never be absolutely guaranteed for determinate rounding.

In another respect determinate rounding is superior: with random-rounding it is possible to determine an f value by averaging a large number of $\hat{f}$ values. For determinate rounding this trick obviously won't work. While relatively unimportant for off line dialogue, this weakness of random-rounding weighs more heavily in on line dialogue.

As long as one insists on zero bias, and wants to exclusively round either up or down to the next multiple of b, for random-rounding there is only one algorithm (25). For determinate rounding there are two alternatives:

(a) Systematic rounding
(b) Hash-rounding (27)

In comparing both possibilities [3] we shall keep in mind that we do not intend to keep the rounding function secret. On line dialogue is more vulnerable, so our most important objective is to obscure as far as possible any trackers (and the identifying conditions for Palme's scheme (27), which are equivalent). We will denote the relevant logical conditions as follows:

T = a tracker (code 10 in the example, section 2.5)
A_1 = a logical condition to be employed alongside with any tracker ((30); code 04 in the example, section 2.5)
P = a pseudotracker (29, 30), which, if employed alongside with A_1, will define a subset of cardinality > 1 (no full identification)

---

[3] For the impulse to inspect more closely, and with an unexpected result, the two rounding schemes, I am indebted to X. Debanne.

Then the following equations hold (see 29, 30)):

$$f(A_1) = f(T \wedge A_1) + 1 \qquad (3.4.1)$$
$$f(A_1) = f(P \wedge A_1) + i \quad (i = 2, 3, 4 \ldots) \qquad (3.4.2)$$

The rounding functions g will be judged by two criteria:

(1) The *visibility* of any trackers formed. Because of (3.4.1) a pair T, A_1 will be visible, provided that

$$\hat{f}(A_1) = g(f(A_1)) \neq f(T \wedge A_1) = g(f(T \wedge A_1))$$

(2) The *ambiguity* residing in any pair of rounded values

$$\left\{\hat{f}(A_1), \hat{f}(X \wedge A_1)\right\}, \text{ where } \hat{f}(A_1) \neq \hat{f}(X \wedge A_1)$$

With $\hat{f}$-values it is not easy for an intruder to find out, whether X is a tracker T, or only a pseudotracker P (29). The discrimination becomes more cumbersome, where some threat monitoring scheme like the one sketched in the previous section sharply cuts the probability that an intruder will "come through" with a fully identifying condition. Finally the job will be the more difficult, the more pairs of numbers $f(A_1)$, $f(P \wedge A_1)$ will, after rounding, result in the same pair of rounded values $\left\{\hat{f}(A_1), \hat{f}(X \wedge A_1)\right\}$, that is, the higher the ambiguity of some pair of f-values.

We will apply the criteria of visibility and ambiguity to systematic rounding and to an extremely hashing function, both with b = 5, and (assuming equidistribution of final digits) with zero bias. The functions, and the trackers remaining visible after rounding, are shown in table 2. The hash function leads to 100%, systematic rounding to 20% visibility only. Generally, systematic rounding with rounding base b will lead to a visibility of $(1/b) \cdot 100\%$; any hashing attempts will increase that percentage.

One still might expect that the confusion introduced by hashing will increase the ambiguity. But this is not the case. $\{m, n\}$ will mean (cf. table 2):

$$\hat{f}'(A_1) = m \quad \text{and} \quad \hat{f}'(X \wedge A_1) = n,$$

while m/n denotes

$$f'(A_1) = m \quad \text{and} \quad f'(X \wedge A_1) = n$$

The comparison will be restricted to the pair $\{5,0\}$. For the remaining pairs the situation either is analogous (e.g., for the pair $\{10, 5\}$), or even more in favor of systematic rounding: pairs like $\{5, 10\}$ or $\{-5, 0\}$ can not result

$f \equiv x \pmod{10}$ $f' = x$	$\hat{f} = g(f) \equiv y \pmod{10}$ Hashing function $\hat{f}' = y$		$\hat{f} = g(f) \equiv y \pmod{10}$ Systematic rounding $\hat{f}' = y$	
-5	-5	+	-5	
-4	0	+	-5	
-3	-5	+	-5	+
-2	0	+	0	
-1	-5	+	0	
0	0	+	0	
1	5	+	0	
2	0	+	0	+
3	5	+	5	
4	0	+	5	
5	5	+	5	
6	10	+	5	
7	5	+	5	+
8	10	+	10	
9	5	+	10	

Table 2:

Systematic rounding and extremely hashing rounding function with rounding base $b = 5$. The + sign denotes a visible tracker; it is attached to the row, whose f' corresponds to $f'(T \wedge A_1)$ – the next lower row then corresponds to the respective $f'(A_1)$, see equation (3.4.1). The + is attached, whereever $\hat{f}'(T \wedge A_1) \neq \hat{f}'(A_1)$.

from systematic rounding, because always $f(A_1) > f(X \wedge A_1)$, but with a hashing function

$$\hat{f}(A_1) < \hat{f}(X \wedge A_1)$$

may well occur. The pair $\{5, 0\}$ then may result from the following pairs of f' – values (compare table 2):

Hashing function					Systematic rounding				
9/4*	9/2*	9/0*	9/–2*	9/–4*	7/2	7/1	7/0	7/–1	7/–2
7/4*	7/2*	7/0*	7/–2*	7/–4*	6/2	6/1	6/0	6/–1	6/–2
$\underline{5/4}$	5/2	5/0	5/–2*	5/–4*	5/2	5/1	5/0	5/–1	5/–2
(3/4)	$\underline{3/2}$	3/0	3/–2*	3/–4*	4/2	4/1	4/0	4/–1	4/–2
(1/4)	(1/2)	$\underline{1/0}$	1/–2*	1/–4*	$\underline{3/2}$	3/1	3/0	3/–1	3/–2

So there are	Hash function	Systematic rounding
(a) Pairs underlined, which constitute a real $f'(A_1) / f'(T \wedge A_1)$ – pair:	3	1

(b) Pairs in brackets ( ), which an intruder may disregard because of the impossibility of $f(A_1) < f(X \wedge A_1)$:	3	none
(c) Pairs with an asterisk *, where testing readily will unmask the pseudotracker, because $\hat{f}'$-values other than 5 or 0 (that is, either 10 or -5 or both) occur between $f'(A_1)$ and $\hat{f}'(X \wedge A_1)$ (for details see (29, 30)):	16	none
(d) Remaining pairs without symbols attached to them, constituting true ambiguities which the intruder carefully has to test for:	3	24

With determinate rounding and $b = 5$ a higher degree of ambiguity than the one achieved by systematic rounding obviously is unattainable. As long, as the rounding function itself is not to be kept secret, systematic rounding, regarding its ability to obscure trackers, is superior to the – intuitively quite attractive – concept of hash rounding.

4. Some long-range implications

War is said to be too earnest to leave to the generals. Mutatis mutandis the same could go for the social and political implications of computing, which are not left to information and computer scientists. But they will be asked to offer some thoughts of their own. In mentioning some of the possible long-range implications of our problem I cannot furnish any definite wisdoms, but just try to provoke some of the thoughts which will be anticipated from us.

To me there seem to be at least two major areas where statistical confidentiality may be expected to pose some uncomfortable questions:

(1) The validity of statistical data
(2) The conflict between the right to know and the right to privacy

The validity and, to some extent, even the availability of statistical data depend to a high degree on the readiness of respondents to cooperate with the investigator. While this may be less critical for pure measurement data – say, blood sugar or body weight – it becomes essential where the respondent has to be asked something, and is supposed to answer truthfully. The overwhelming and sometimes – regarding validity of results – devastating influence of the experimenter-subject relation on the results of psychological tests and experiments as well as on the contents of questionnaires is well known to social psychologists ((22), especially chapters 15–19). The experiments of Asch (2, 3) and his successors (review in (1), pp. 16–23) have demonstrated beyond doubt that, given sufficient social pressure, a subject will tell, not what he thinks or clearly perceives, but what he thinks to be wanted to tell. The experiments also have shown that confidence in anonymity of the answers will partially cancel the effects of social pressure.

Experiments of this kind do not constitute a paradigma which may be simply transferred to all situations of collecting statistical data. But they certainly don't contradict the fears repeatedly expressed (7, 9, 16) that neglecting statistical confidentiality in the long run might result in a significant loss of availability and validity of statistical information. The risks described to you are practically meaningless for small samples of records, whose composition is kept confidential. But you might run into difficulties when trying to explain this little difference to a respondent once he has become distrustful.

The conflict between right to know (or freedom of information) and right to privacy (or, in our case, the respondents' right to anonymity) may well foster some paradoxial implications.

Privacy is a prerequisite for any open society. "Just as a social balance favoring disclosure and surveillance over privacy is a functional necessity for totalitarian systems, so a balance that ensures strong citadels of individual and group privacy and limits both disclosure and surveillance is a prerequisite for liberal democratic societies." (Westin (39), p. 24).

But freedom of information, too, is a prerequisite for an open society. "The democratic society relies on publicity as a control over government" ((39), l.c.) – and over any powerful organizations, we might add. Citizens, journalists, scientists, and their organizations, claim free access to information sources, and with good reasons. Data banks and teleprocessing offer the prospect that we all become less dependant on what the Lords of Things, who control, among others, the dissemination of statistical information, are able or willing to tell us: one day we might go to our library, sit down at some terminal, inspect the original data, and form our own independent judgment.

There is no guarantee that the balance of knowledge will tip that way (12). It seems a good bet that we sooner or later will witness attempts to look information sources in the name of privacy and an open society. Confidentiality risks will be pretended where thwarting information claims is meant – it's just too tempting an excuse. Information and computer scientists might well find it worthwile to rack their brains, lest this excuse work.

Literature

(1) Aronson, E.: The social animal. Freeman, San Francisco 1972.

(2) Asch, S.E.: Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychol. Monographs Gen. Applied 70, Nr. 9 (1956).

(3) Asch, S.E.: Opinions and social pressure. Scientif. Amer. 193, 5 (Nov. 1955), 31–35.

(4) Assakul, K. and C.H. Proctor: Testing independence in two-way contingency tables with data subject to misclassification. Psychometrika 32 (1967), 67–76.

(5) Astin, A.W. and R.F. Boruch: A "link" system for assuring confidentiality of research data in longitudinal studies. Amer. Educat. Res. J. 7, (1970), 615–624.

(6) Baran, P.: On distributed communications: IX. Security, secrecy, and tamper-free considerations. Memorandum RM-3765-PR, The Rand corporation, Santa Monica, Calif., August 1964.

(7) Boruch, R.F.: Security of information processing – implications from social research. AFIPS Conf. Proc. 41, I (1972), 425–433.

(8) Boruch, R.F.: Strategies for eliciting and merging confidential social research data. Policy Sci. 3, (1972), 275–297.

(9) Boruch, R.F. and G. Endruweit: Mathematische Methoden zur Sicherung der Vertraulichkeit und Anonymität von Forschungsdaten. Zschr. Soziol. 2 (1973), 227–238.

(10) Dowling, T.A. and R.H. Shachtman: On the relative efficiency of randomized response models. J. Amer. Statist. Ass. 70 (1975), 84 -87.

(11) Eimeren, W.v., H.-K. Selbmann and K. Überla: Modell einer allgemeinen Vorsorgeuntersuchung im Jahr 1969/70. Schlußbericht. W.E. Weinmann Druckerei GmbH, Bonlanden b. Stuttgart 1972.

(12) Fano, R.M.: On the social role of computer communications. Proc. IEEE 60 (1972), 1249–1253.

(13) Farr, M.A.L., B. Chadwick and K.K. Wong: Security for computer systems. National Computing Centre, Manchester 1973.

(14) Feige, E.L. and H.W. Watts: Protection of privacy through microaggregation. In R.L. Bisco (ed.): Data bases, computers, and the social sciences, 261–272; Wiley, New York 1970.

(15) Feige, E.L. and H.W. Watts: An investigation of the consequences of partial aggregation of micro-economic data. Econometrica 2 (1972), 343–360.

(16) Fellegi, I.P.: On the question of statistical confidentiality. J. Amer. Statist. Ass. 67 (1972), 7–18.

(17) Fellegi, I.P. and J.L. Phillips: Statistical confidentiality: some theory and applications to data dissemination. Ann. Econ. Soc. Measurement 3 (1974), 399–409.

(18) Hansen, M.H.: Insuring confidentiality of individual records in data storage and retrieval for statistical purposes. AFIPS Conf. Proc. 39 (1971), 579–585.

(19) Hentschel, B., H. Gliss, R. Bayer and R. Dierstein: Datenschutzfibel unter besonderer Berücksichtigung des Personalwesens. Bachem, Köln 1974.

(20) Hoffman, L.J. and W.F. Miller: Getting a personal dossier from a statistical data bank. Datamation 16, 5 (May 1970), 74–75.

(21) Jacobs, G.: Die Unwirksamkeit der Anonymisierung von Individualdaten – dargestellt am Beispiel der Amtlichen Studentenstatistik. Öffentliche Verwaltung und Datenverarbeitung 3 (1973), 258–261.

(22) Jourard, S.M.: Self-disclosure. An experimental analysis of the transparent self. Wiley-Interscience, New York etc. 1971.

(23) Liu, P.T., L.P. Chow and W.H. Mosley: Use of the randomized response technique with a new randomizing device. J. Amer. Statist. Ass. 70 (1975), 329–332.

(24) Müller, P.J.: Datenschutz und Sicherung der Individualdaten der empirischen Sozialforschung. Datenverarbeitung in Steuer, Wirtschaft und Recht 3 (1974), 2–11.

(25) Nargundkar, M.S. and W. Saveland: Random-rounding: a means of preventing disclosure of information about individual respondents in aggregate data. Statistics Canada, Ottawa July 1972.

(26) Nargundkar, M.S. and W. Saveland: Random-rounding to prevent statistical disclosures. Amer. Statist. Ass. Proc., Soc. Statist. Section (1972), 382–385.

(27) Palme, J.: Software security. Datamation 20, 1 (Jan. 1974), 51–55.

(28) Reed, I.S.: Information theory and privacy in data banks. AFPIS Conf. Proc. 42 (1973), 581–587.

(29) Schlörer, J.: Schnüffeltechniken und Schutzmaßnahmen bei statistischen Datenbank-Informationssystemen mit Dialogauswertung. Materialien Nr. 29 der Abteilung für Medizinische Statistik, Dokumentation und Datenverarbeitung der Universität Ulm, [illegible].

(30) Schlörer, J.: Identification and retrieval of personal records from a statistical data bank. Meth. Inform. Med. 14 (1975), 7–13.

(31) Schlörer, J.: On statistical confidentiality: Minimum query language requirements for tracker building. Materialien Nr. 33 der Abteilung für Medizinische Statistik, Dokumentation und Datenverarbeitung der Universität Ulm, 1975.

(32) Schlörer, J.: Zum Statistikgeheimnis: Risiken und Schutz statistischer Datenbanken. Datenv. im Recht (DVR) 5 (1976), 203–248.

(33) Schlörer, J.: On statistical confidentiality: a threat-monitoring scheme for on line dialogue. Meth. Inform. Med. 15 (1976), 36–42.

(34) Selbmann, H.-K.: Ein Datenbanksystem zur Auswertung statistischer Datenbestände. Materialien Nr. 15 der Abteilung für Medizinische Statistik, Dokumentation und Datenverarbeitung der Universität Ulm, 2. Auflage, Febr. 1973.

(35) Selbmann, H.-K.: Bitstring processing for statistical evaluation of large volumes of medical data. Meth. Inform. Med. 13 (1974), 61–64.

(36) Turn, R.: Privacy and security in personal information databank systems. The Rand Corporation, Santa Monica, Calif., R-1044-NSF, March 1974.

(37) Turn, R. and N.Z. Shapiro: Privacy and security in databank systems – measures of effectiveness, costs, and protector-intruder interactions. AFIPS Conf. Proc. 41, I (1972), 435–444.

(38) Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. J. Amer. Statist. Ass. 60 (1965), 63–69.

(39) Westin, A.F.: Privacy and freedom. Atheneum, New York 1970 (6th printing).

Social-Information-System in the Federal Republic of Germany

The Data-Base of the Federal Minister of Labour and Social Affairs

Herbert Schmidt, Bonn

1. Socio-political need for the creation of a social data benk

Since the beginning of the industrialization process with all its repercussions on working life and on the social correlations, social policy has become more important in all countries. The effects of technical, organizational and social changes on the structure of our society in general have become more complex and intricate. The pressure for government initiative in this field has become stronger and stronger. The multitude of statutory measures to safeguard the social security of the individual and the population as a whole in a liberal state based on the rule of law, makes it imperative to create possibilities for a greater transparency in this field of developments partly diffuse and yet interdependent.

The increased importance of social policy for society in the Federal Republic of Germany is demonstrated most impressively by quoting the *rate of social benefits*, i.e. the calculated value of the ratio between the annual amount of social benefits and the gross national product for the same year. In 1975 this rate amounted to 27.1%. This means that according to the data on social benefits covered by the social budget 1973 more than one quarter of the gross national product was spent for social purposes in the Federal Republic of Germany. Social benefits as defined in the social budget comprise preventive, curative, and integrating measures as well as cash benefits plus allowances in the form of tay exemptions and price reductions fixed by law which are granted if certain social prerequisites are fulfilled by special statutory institutions, by federal, local or regional governments and by other undertakings by virtue of legal provisions, by-laws, collective agreements or voluntary arrangements (1). It must be expected that in future the social benefits will increase more rapidly than the gross national product. Thus the rate of social benefits will presumably amount to more than 29% by 1978. In absolute figures this means that in the Federal Republic of Germany in 1974 about 287 000 million DM will have been spent on direct and indirect social benefits. For 1978 social benefits are expected to amount to about 430 000 million DM.

The intensity of the government's activity in the field of social policy and the social benefits resulting therefrom primarily depend on the following factors:

- the state of social legislation,
- the age structure of the population,
- the state of health of the persons covered by social security and the demand for the benefits offered,
- the development of productivity, prices, wages and employment.

These decisive factors in turn depend on the manifold dynamic processes influencing our society.

In the interest of planning, implementation, and control of social measures, as well as in the interest of their potential or actual impact on society, it is necessary to dispose of a comprehensive information system for statistical data, projections, estimation, and simulation methods as well as corresponding programme packages.

2. Rationalization requirements and possibilities for establishing a combination system as a basis for the creation of a social data bank

The extend and the complexity of the tasks to be carried out in the interest of social security and humanization of working life require an efficient administrative machinery and significant assessment criteria for those responsible and interested in all spheres of social life.

Without a computerized compatible data collection, data combination, and data analysis a purposeful assessment of the actual state and the target state of activities in the field of social security is no longer feasible in the long run. Data requirements can be inferred from the above mentioned global component, requirements which are necessary in order to take action in the framework of a systematic and farsighted social policy. Moreover the growing importance of social policy, as outlined above, for the modern industrial society and its interdependence with other fields of private and state activities emphasize the need for the establishment of integrated information combination systems in the interest of a greater transparency, a careful planning of social policy and the subsequent responsible utilization of funds in the interest of social security.

The great need for data to be used by government and parliament, by the social partners, by science, industry, and the population induced the Federal Ministry of Labour and Social Affairs to develop and establish a social information system of its own (social data bank) starting from 1971. The motivations underlying this measure were above all the need for statistics of persons employed which were repeatedly requested by the Federal Government's Council of Experts, and the urgent necessity of a greater transparency of social insurance.

The attainment of this aim was facilitated by the increasing computerization in the social insurance institutions. Since the general public has become aware of the need for private and state activities in the field of social security the collection of "mass data" in the field of social policy has become a matter of course 1). This system of social security in the narrower sense of the word comprises the Federal Institute for Employment with its regional and local labour offices, about 1 600 sickness insurance funds as well as about 30 panel doctors' and panel dentists' association,, approximately 100 accident insurance institutions and per Land one regional assistance service for war victims as well as the factory inspectorate.

In the broader sense of the word social security data arise where human beings are integrated into the work process: working places in industry, in the different organizations and associations as well as in authorities. Data relevant to the different branches of economy accrue in firms in the fields of manpower management, training and social services and give a picture of the social structure and the social security of the staff of the respective establishment and, in their aggregate form, contain important structural information relating to the overall economy. Such data are of particular importance with regard to the need for the state to become active or to remain passive in certain fields of social policy. A great number of data from the most different sectors of our economy flows to the social security institutions in the narrower sense of the word.

With the increased utilization of data processing in the areas of social security described above the possibility of developing an independent information system in the interest of improved information for the citizen and the state by means of a data combination system emerged. Particularly the "Regulation of the Collection of Data for the Social Insurance Institutions and for the Federal Institute for Employment" (Data Collection Regulation – DEVO –) dated 24/12/1972 (Bundesgesetzblatt I, p. 2159) as well as the "Regulation on Data Transfer by Machine Usable Data Carriers in the Field of Social Insurance and of the Federal Institute for Employment" (Data Transfer Regulation –DÜVO –) dated 18/11/1972 (Bundesgesetzblatt I, p. 2482) provided the

---

1) The abundance of individual data accumulating in the field of social security is borne out by the fact that already in 1912 when the punchcard system was developed, a social insurance system for the Ruhr area was the first to introduce the novel equipment in Germany. This necessity to implement new and modern methods in the interest of a rationalization and increased efficiency of the public service continues to exist in the Federal Republic of Germany in the early age of electronic data processing. In 1956 the Federal Insurance Institution for Salaried Employees in Berlin was the first agency to introduce a data processing system into public administration in the Federal Republic of Germany.

basis for the establishment of a social data bank at the Federal Ministry of Labour and Social Affairs.

Since the beginning of 1974 data from 170 information systems in the field of social security as well as from far more than 1000 DP-based information systems of firms and plant-operated sickness insurance funds are aggregated to data files in the Federal Ministry of Labour and Social Affairs which in this form are unique in the international framework.

3. Technical and organizational bases of the social data bank

The storage and processing of data in the social data banks is done by Unit 45 of the Siemens-system 4004. It was set up in October 1973 at the Federal Ministry of Labour and Social Affairs in Bonn. The processing of formatted data is carried out by the SESAM-system. The processing of non-formatted data is carried out by the GOLEM-system.

In view of the volume of data to be electronically processed at the Federal Ministry of Labour and Social Affairs, the demands on the unit almost exceed its capacity. During certain periods it is necessary to work up to three shifts a day. Moreover, the data retrieval system SESAM is overtasked in a number of areas by excessive demands. Only relatively simple retrieval operations can be carried out, and these still consume a considerable amount of time. It is a matter of course that increasing tasks and more sophisticated evaluation programmes, complicated projection methods and simulation models require more effective equipment in respect to comfort and capacity. It is therefore intended to install a Unidata 7-750 Time-Sharing Unit.

When the social data bank was established, the following principles were applied (2, 3):

(1) the traditional reporting system in the social security field is to serve as a guideline for the establishment of the social data bank;

(2) agencies which collect social security data at the source must supply their data as a rule in machine-readable form; the application of this principle is to minimize multiple storing, manual work and other administrative procedures for all institutions concerned;

(3) standardization and the possibility of data combination must be ensured by uniformity of documentation and print layouts; forms and ways of data flow are determined in consultation with all partners associated in the data combination system;

(4) the social data bank is to store a large range of data covering the individual industries and members of the labour force. Only anonymous data are fed into the social data bank, thus ensuring manifold possibilities of evaluations and aggregations. Losses of information as they happen sometimes in conventional statistical reporting and aggregation procedures are to be avoided;

(5) in the interest of a maximum degree of up-to-dateness of the data stored at the social data bank, all efforts are made to shorten the reporting channels as much as possible;

(6) storage and compilation of data at the social data bank are to ensure the automatic production of statistics, tables and reports in a form ready for publication:

(7) by the development of an information service for the technical divisions and sections of the Federal Ministry of Labour, the data needed by them are concentrated at one place; this service is to replace the collection of data by the individual technical sections for their particular fields of interest;

(8) the data bank system is to ensure, by means of data display devices, the interactive communication with the computer to arrive at a rapid information retrieval;

(9) the social data bank is conceived as an open-loop system; it is intended to make this bank including its extension available to a broad group of users.

4. Contents of the social data bank

The social data bank comprises in particular data from the following sectors of social security and labour law:

a) statutory sickness insurance
b) statutory accident insurance
c) statutory pension insurance
d) employment services
e) assistance to war victims
f) collective agreements
g) factory inspectorate.

As an example there will be given further information about the first individual sector of the social data bank: the statutory sickness insurance.

Statutory sickness insurance

The Federal Ministry of Labour and Social Affairs receives aggregate data only for the field of statutory sickness insurance which covers about 55 million insured persons (including family dependants); these data come in separately from the almost 1 600 sickness insurance funds in Federal Republic of Germany. In detail the social data bank stores the following information:

- membership, number of sick persons and rates of contributions
- age structure of the members
- breakdown of the members by Länder
- data on the by-laws and personnel of the sickness insurance funds (about 78 000 persons)
- contingencies
- preventive measures
- profit and loss account under the Act on the continued pay of wages in case of sickness
- benefits in the framework of welfare to convalescent persons and other welfare measures
- statement of accounts (receipts, expenditure, capital assets)
- the receipts and expenditure of the overall sickness insurance system at present amount to over 50 000 million DM a year
- services commissioned by other insurance institutions
- data on the number of members and pensioners for equalization purposes
- receipts from contributions, expenditure for benefits, individual assets (per quarter).

5. Supplementation of the data file

The efforts of the Federal Ministry of Labour to supplement and complete the relevant social data go partly beyond the province of the labour ministry. In conformity with the responsibility of the Federal Labour Minister to establish the social budget for the Federal Republic of Germany, the collection of social data or the derivation of social indicators respectively, the compilation of data regarding measures of training, further training and retraining within the framework of statistics of persons employed and the promotion of vocational training under the Employment Promotion Act through the Federal Institute for Employment, gerontologic data regarding older workers or workers who have retired from working life and pensioners etc., the social data system of the Federal Ministry of Labour and Social Affairs plays a central part in the establishment of the envisaged federal information system. This also applies to general information and the evaluation for scientific purposes within the framework of a specialized information scheme "social sciences", the development of which was initiated by the action programme "Information and Documentation" launched by the Federal Government.

In view of the demand of the general public, in particular the active population, the trade unions, the employers organizations, industry, science and public agencies, to make political and structural factors and changes in the employment conditions of today more transparent and to accelerate the dissemination of new findings regarding the humanization of working life, the Federal Minister of Labour and Social Affairs intends to complete his central (in particular numerical) information system in order to lay the foundation for a technical information system covering specific manpower and social security problems by the compilation of a pertinent literature and research documentation. Such a compilation of texts would also offer the possibility of storing and retrieving social and labour law standards, court judgements and evaluations of literature.

6. Documentation on labour law and social law

The increasing lack of transparency of social legislation and adjacent fields as well as the resulting jurisdiction makes it necessary to utilize the possibilities of modern data processing with a view to improving the transparency of the applicable legislations. The Federal Minister of Labour and Social Affairs, in close cooperation with the central federal courts, the Federal Social Court and the Federal Labour Court, has examined the possibility of establishing a relevant legal data bank. As a first step, the data bank on the administration of social law was established in collaboration with the Federal Minister of Justice. This data bank with 20 terminals will take up operations already in the course of 1975 (4).

The development of the documentation of social law was facilitated by the preliminary work of the Federal Social Court. Over a period of more than 20 years this court had already developed a reliable register. The data bank on social law will contain legal standards, judgements and relevant literature. Since its establishment in 1954, the Federal Social Court (Bundessozialgericht – BSG) has evaluated approximately 225 000 documents regarding standard-setting jurisdiction with guidelines or brief summaries all of which are of importance for the administration of social law. The storage increases by approximately 12 000 documents annually.

With the help of the data bank system GOLEM 2 developed by the firm Siemens AG out of GOLEM 1, together with several supplementary programmes, this data bank will test the effectiveness of such a system before we start establishing a comprehensive legal information system. In the course of this work, data collection catalogue, data compilation designs with instructions and data compilation forms were elaborated. At present approximately 16 000 judgements passed by social courts and listed in the file of the Federal Social Court were put into the data bank on social law. The data are composed of approximately 6 000 decisions passed by the Federal Social Court, 9 000 decisions passed by the Regional Social Courts and 1 000 decisions passed by local social courts. Because the Regional Social Courts utilize the data compilation forms elaborated by the Federal Social Court, the necessary compatibility of data is ensured. Also relevant court decisions of other courts are to be made accessible. 9 insurance institutions and associations have already declared their willingness to contribute to the further expansion of the data bank on social law by editing their documentation in conformity with the prescribed data compilation forms. This procedure will make it possible to evaluate and store periodicals, directives, by-laws, decisions, guidelines etc.

The group of users of the data bank on social law will be composed primarily of representatives of the executive and the legislative powers, the social courts, the social insurance institutions and their associations, the trade unions, the employers' associations, the undertakings, lawyers and scientists. The services to be provided by the data bank may take the following three forms:

(1) the retrieval of information by direct approach with the help of data display devices
(2) information by cards or microcards on the special field of interest of the user
(3) supply of laws, judgements or literature by data bank information printing. In this way the user will receive a microcard file.

This data bank will offer a variety of advantages for all persons concerned. Such advantages consist in time saving, more rapid information, overall survey and elimination of leafing through documents or establishing files. Experts expect that in 1978 at the latest the data bank on social law will have attained such a degree of perfection that non-users of this information system who persist in conventional documentation methods will, because of inevitable gaps in their information, find themselves at a disadvantage. The success of the initiative to set up the data bank on social law has stimulated a corresponding documentation process in the field of labour law. When developing the data bank on labour law, account will be taken of the requirements of compatibility in order to make it accessible to a minimum number of users.

7. Simulation models in the framework of government activities

It is possible to calculate "play through" and simulate on the basis of simulation models alternative future developments or circumstances. By simulation models we understand quantitative mathematical models based on a series of hypotheses which are in a given way related to each other. On the basis of empirical data or estimated values mathematical methods are used to simulate projections of the development of variables and trunk groups of variables. Because of many hardly calculable factors descriptions of past or projections of future developments or stages of development in the form of mathematical functions or models could not be made. The lack of transparency of social, economic, and industrial correlations rendered it more difficult or even impossible to reach reliable solutions. The increasing development of automated data processing, the further development of theoretical findings derived from models and their conversion by modern data processing offered new opportunities. The possibility of tapping an up to now unknown empirical data reserve was a further precondition for the improvement of a practical, future-oriented time accelerated x-ray of conceivable development processes. Prospective developments can be appraised, the behaviour and reaction of the protagonists – in government, industry or private households – can be more easily changed and adapted to requirements.

Simulation models for complex problems are highly complicated procedures and require a broad data basis. In the field of social security this data basis is being accumulated by the mass data bank. As far as such projections were carried out at all many mistakes attributable to the simpler projection procedures of the past can be excluded with the growing empirical and theoretical insight into social development processes. Thanks to a greater refinement of the simulation models and procedures we will be able to make such projections in the future. Moreover it will become possible within a relatively short time to change certain hypotheses for given government activities and changing patterns of behaviour of those affected (population, undertakings of different sectors of economy, competition, income developments of the households, level of employment etc.) and thus to obtain information on alternatives for achieving the social aims pursued by the state.

In principle it can be said that the increasing capacity of EDP-based data systems, the high storage capacity and their high computing speed make it possible not only to achieve accurate results on a fixed data base. Moreover due to this technical-scientific progress also more complex probability calculations can be carried out, operations which with the conventional procedure based on manual auxiliary equipment had only achieved inadequate results. For the manifold hypotheses which, in the course of the examination of socio-economic development processes, could be established but not tested for their reliability an analysis can nowadays at least be attempted. The prerequisite is the possibility of examining given circumstances or hypotheses as to their relevancy to probable developments. If the result of this examination is negative such hypotheses can be eliminated from further considerations 2); the complexity of a projection or simulation mode is so being reduced and the probability degree of the forecast resulting from such calculations increases.

With the increasing possibility of verifying model computations and the degree of reality compared with the actual course of development it might also become possible to counteract a "old evil" in the field of government activity: Measures according to the "sprinkling" principle. Many government programmes, laws, assistance measures etc. did not at all or not fully achieve their aim or their full degree of efficiency just because the socio-economic development process took another course than the one expected by the institutions launching such measures. Accordingly various government measures were necessarily broader than would have been required in the interest of the cause itself. Even nowadays the scope of a measure must often necessarily be broader, this means designed in more general terms because certain target groups, target sectors or target fields could possibly not be reached otherwise. On the basis of a sound documentation for the decision-making process such global controls could in future perhaps be reduced, public funds utilized more efficiently, economic and social progress planned more realistically and be steered with more refined instruments of government policy 3) for the benefit of the entire society.

8. Simulation computations for the socio-political decision-making process

a) Model-categories

Tentative projections and simulation models for the assessment of future developments resulting from socio-political measures or socio-economically relevant

---

2) In this connection Giersch points to the fact that in the extreme, i.e. "if information on the future is complete all hypotheses but one will be eliminated; this hypothesis will then reach the probability degree in full and thus be given the degrees. Thus the extremes of complete ignorance and complete information are labelled".
Giersch, Herbert: Allgemeine Wirtschaftspolitik. Wiesbaden 1960, pp. 41.

3) In the interest of a better basis for draft legislation simulation results have meanwhile been laid down in the explanatory report accompanying bills. There is even a statutory requirement to this effect with regard to the submission of the pension adjustment report; with regard to the social report the structural analyses and the development projections were explicitly demanded by a declaration of government. Further reference is given in the text.

events have often been carried out on the basis of quantitative mathematical models. In this context for example the calculation of the financial effects of the 1957 pension reform as well as the computation of prognostic values in connection with the establishment of the first social budget in 1968 or various considerations about the demand for manpower on the labour market have to be emphasized.

The establishment of the social data bank at the Federal Ministry of Labour and Social Affairs set new priorities and opened up new perspectives. The hard efforts for the collection of additional data, the improvement of the compatibility of data, the establishment of a data combination system and other standardization efforts led to a qualitative improvement of the basic data for sociopolitical decisions. This in turn in an improvement of the models used in the Federal Ministery of Labour which again had a positive effect on the accuracy of projection values. Moreover, the advantages of electronics became apparent in the acceleration of the calculation process as well as with regard to the criteria to be taken into account and the flexibility in case of new or alternative computations.

In the Federal Ministry of Labour a clear distinction is made between two types of simulation models (5):

(1) *Models with a general monitoring function*
This model category comprises models dealing with the actual state of affairs and with developments to be projected on the basis of the ex-post data. The indicators obtained by these models can be considered as an early warning system in order to induce the government to take action or to propose solutions to the legislator. For example this category of model computation comprises the tentative projections in the framework of the social budget, the annual pension insurance projections as well as the envisaged projections in the field of statutory sickness insurance.

(2) *Models for analyzing and simulating complex correlations*
This type of models includes for example the model for the sickness insurance for pensioners, the second model for the Amendment to the Sickness Insurance Act, the simulation of the pension reform legislation 1972 as well as the structural model on statutory sickness insurance. In these models the explanation of *cause and effect* of political measures is the major object of the analysis.

In contrast to the first type of models (1) with their more *informative character* which is based on the given legal situation, in the second type of models (2) the legal situation is considered a variable. This category must in any case contain statements on the correlations between (instrument) variables (to be influenced politically) and politically relevant states of affairs (target variables).

Geißler (5) subdivides the above mentioned type of models into two categories: *projection models* and *structural models*. The first category includes the system projections, for example the pension insurance projections and the projection model for the sickness insurance of pensioners, as well as differentiated and modified models for the trend extrapolation. The latter category includes the projection models for the statutory pension insurance as well as for the social budget. Most trend extrapolation models can be considered as conventional simulators whereas the system projection models as well as the structural models are in a particular need of a computer based system because of their complexity and efficiency analysis.

A few comments should be made on the various simulation models in the field of social security:

b) Projection and structural model for the statutory sickness insurance scheme

In recent years, the problems of statutory sickness insurance created by the cost explosion in the field of medical care have come to the foreground of social policy and public health discussions in the Federal Republic of Germany. Although the rate of contributions was stepped up from 8.2% to 9.5% between 1970 and 1974, an end of contribution increases as a consequence of the growing financial burden on sickness insurance institutions cannot yet be predicted. In order to arrive at a better analysis of causes and to test possible instruments for controlling the further development in this field, a simulation model for medium-term developments of receipts and expenditure within the statutory sickness insurance scheme was set up at the Federal Ministry of Labour. This differentiated and modified trend extrapolation model can only function with the help of computerized operations.

On the basis of the data on benefits and financing provided by approximately 1600 sickness insurance funds in the Federal Republic of Germany, the future complex structural model for statutory sickness insurance will be able
- to demonstrate the present contribution differentials and
- to simulate the effects of potential measures on the individual sickness insurance fund or sickness insurance structure, respectively.

In this connection reference should be made to the existence of two simulation models for the implications of the Act governing Sickness Insurance for Farmers (1972) and to the simulation model regarding sickness insurance for pensioners. The first-mentioned models permit to predict the probable number of persons willing to join the envisaged agricultural sickness insurance funds as well as the age structure of this membership. Moreover, these models have made it possible to carry out calculations regarding contribution scales.

The development of the simulation model regarding sickness insurance for pensioners became necessary bacause it was impossible to reply to the political and technical questions concerning the financial burden and the distribution problems involved and to test different political alternatives and proposals the conventional methods of analysis. The direct cause for the creation of this simulation model was the sky-rocketing expenditure of the approximately 1600 sickness insurance funds as a consequence of the rise of prices and costs in the medical field. The burden involved varied from one sickness insurance fund to the other because of the difference in the structure of membership and in particular because of the proportion of pensioners affiliated to a given sickness insurance institution. Moreover, the expenditure for pensioners is not covered by the contributions paid by the pension insurance institution to the statutory sickness insurance scheme. Since the costs for medical care rise more quickly than the pensions, the contributions paid by the pension insurance scheme – amounting to a total of 11% of the pension volume – covered less and less of the actual expenditure. The financial problem resulting from this situation for the sickness insurance institutions is evidenced by the fact that in 1968, 80% of the expenditure for pensioners borne by the sickness insurance funds were covered by contributions as against somewhat less than 60% in 1973. In view of the variety of problems resulting therefrom as well as in view of the competitive situation between the different types of sickness insurance funds and the manifold political obstacles, the computerized simulation model offered the possibility of neutralizing the discussion and rationalizing decisions with the objective to create equitable statutory regulations. Contrary to the past, the utilization of computers enabled us – despite diverging proposals and hypotheses – to reduce emotions, to disprove the conventional approaches and to develop a new contribution formula based on the actual expenditure which ensured an absolute equality of treatment 4). This "simulation result" is reflected in a ministerial draft of the Fourth Amendment to the Sickness Insurance Act (1974) which is at present being discussed by the parliamentary bodies.

---

4) Geißler calls the results of this simulation model constructed by him as "somewhat sensational". He rightly points out that by means of simulation, decisions are no longer taken "blindfold", in other words unable to overlook the actual effects of a measure because of the complexity of its impact. Cf. Ulrich Geißler: op. cit.

From a data processing point of view, it must be emphasized that this model offers the possibility of interactive communication. Approximately 20 000 data supplied by the sickness insurance funds on punch cards were fed into the model having recourse for the first time also to graphic methods. In this way it became possible to present the results of simulation by means of data display devices as well as by plotters.

c) Need for projections within the framework of the social budget

The social budget first published in 1968 has the objective to provide a modern and transparent picture of social policy. This so-called "policy with pockets of glass" (6) is above all intended to improve the transparency with regard to the interdependence of social policy, economic policy and fiscal policy. Moreover it is to contribute to a better understanding of the tasks and the course to be pursued by a social policy embracing the society as a whole. The well founded analytical methods succeeded in overcoming and eliminating the prevalently emotional approach that the governmental social policy was based on ideals of social romanticism or social utopia and, because of its imponderability, was to be considered as a romantic and utopic approach in itself.

The social budget [5] (7) is presented annually and projects the future development of social expenditure and its financial implications for a four-year period on the basis of a medium-term economic forecast made by Federal Government. Thanks to the further elaboration of the concept of the social budget, it has meanwhile become possible to accurately delineate developments in the various sectors of social security and to compute the rate of social expenditure, i.e. the ratio the overall social expenditure bears to the gross national product of the Federal Republic of Germany.

The growing volume of data and the extension of the social budget led to increasing demands for the computation of new items and the identification of interdependencies with other sectors of social security and with economic trends. The number of problems and tasks involved, the need for an insight into the interdependence of developments, the inclusion of relevant data etc. made it necessary to elaborate more flexible methods for short-term adjustments of the models to actual benchmark data, as well as to sectoral and social trends. In the long run, an accurate gearing to the general economic projection is only possible with the help of electronic data processing. Since 1972, this complicated mathematical process is carried out by programmed batch operations at the electronic data processing system attached to the Federal Ministry of Labour.

---

5) Last presented in 1974 in parliamentary document 7/2853 dated 27-11-74.

9. Further development of the existing programme system for socio-economic model computations

While we have referred to the reality of a number of simulations and projections, we must not delude ourselves as to the obstacles we have to overcome before this objective will have been fully attained. This is particularly true for the computerized development of various models with the objective to design complex model systems by the combination of various models. The example of the social data bank has illustrated that the problem of bottlenecks due to a lack of data for the development and implementation of planning models in the political and administrative field will diminish pari passu with the accumulation of data. In the course of the development of a model we may be faced with problems the solution of which requires the disposal of data that have not been collected so far; on the other hand, the existing data volume may compel us – in the interest of its utilization – to develop appropriate analysis models which enables us to make society-oriented forecasts.

An enquiry carried out by the Institute for Planning and Decision-making Systems operated by the Society for Mathematics and Data Processing showed that computer-oriented planning models in the political field are, as a rule, based on the implementation of a specific solution for a pending problem by means of a problem-oriented language. Time and again efficient planning is hampered, when planning experts are compelled to solve data processing and even programming problems over and above their actual task of analysis and model design. Apart from the fact, that they are thus prevented from concentrating all their efforts on the study of the actual problem, this situation has in the past often led to programme solutions which remained considerably below the optimum level.

On the other hand, EDP-experts are often faced with the task to develop analysis models without sufficient technical assistance, so that such models are subsequently viewed with much scepticism by the actual users. This scepticism is mainly due to the disparity between reality and model as a consequence of an implementation which is primarily geared to the achievement of optimum results. Ulrich makes reference to the limits of model design and simulation in pointing out that "there is a danger for the model to lead a life of its own which is finally only understood by a few experts." To avoid this situation, more computations will be necessary, which is only possible with EDP and specific programme-oriented languages, for instance for simulations purposes (8).

The development of a universal and compatible programme system for the flexible design and execution of projection and simulation models could render valuable services in the solution of this conflict. The Federal Minister of Labour and Social Affairs intends, therefore, to examine, in collaboration with the Society for Mathematics and Data Processing, by means of some pilot projects, the extent to which the development and testing of new software for the design, management, and evaluation of formalized planning models can contribute towards the preparation of decisions in the governmental field. The functions of this programme system shall permit to open up a large range of computerized model calculations. It is further envisaged to elaborate a monitoring language which, besides all, combination and processing of selected partial models, will also take over the linkage of data and model banks for batch as well as for conversational mode. It is intented to develop system functions to assist the solution of the following tasks:

model design
model linkage
model management
model documentation
control of model calling
control of model scheduling
variant construction and management (model structure and model parameter)
editing, management and analysis of the results, having special regard to alternative computations (model structure and model parameter variante).

The results of this development work will ultimately increase the usability of the social data bank. Also non-EDP specialists will, in this way, be able to solve certain questions by means of planning models which are designed and controlled with the help of this model base system. Moreover, maximum compatibility of systems is to be achieved by programming as many system functions in FORTRAN IV as possible. It is expected that the model base system will be capable of supplying or storing the data necessary for the execution of a model. The respective data structures and methods of access will determine the development of appropriate system components. Existing data bank or data management systems are to serve as far as possible as bases for this model bank system, if need be, implemented by interfaces.

The models are to be developed in conformity with the modular concept, in other words many partial models will form one overall model which is the prerequisite for examinations in very heterogenous sectors, such as the simulation of alternative social, economic-technological or other developments. The modular principle will make it possible to gradually include new problems into the complex overall model. Apart from the increase of flexibility, this will also contribute towards a qualitative improvement of projection or simulation results. In pursuance of this objective it may become possible to assign another task to the model base system, namely the consistence check in case of linkage of several partial models.

The advantages of such a user-oriented model base system are the following:

a) Improvement of equality of results of model computations by the flexible inclusion of complex interrelations (linkage possibilities);
b) intensification of the collaboration between the different institutions and their joint data basis;
c) the development of complementary relations to other models; socio-economic systems can, by way of combination, be made accessible to other institutions;
d) elimination of technical and organizational obstacles hampering the implementation of planning models, for instance by the integration of existing partial models;
e) exploration of new avenues for the application of EDP-methods;
f) relief of planning experts from the burden of studying comprehensive programming and operating systems; utilization of the model base system for interactive communications.

The limitations of existing model software produced by large computer manufacturers come to light whenever comprehensive models with a great variety of equations and hypotheses are to be executed. Specific simulation laguages are of some help in this connection but their possibilities are restricted and their application is not sufficiently known. Method banks which also take over management and documentation as well as the support of call and execution of the various method programmes are simply overtasked when they are expected to apply also "exotic" methods which are necessary for econometric models in particular. The farthest reaching development in this direction is the TROLL-system (9, 10, 11) which was developed within the framework of the MIT-prognosis project. It offers standardized input, output, and data management routines for model design as well as a specific monitoring language for execution and simulation of a model and does not require the user to have any special knowledge of the operating system.

The concepts enumerated in this paper are only partial solutions for a model base system which will support the flexible management, processing, and linkage of models and data. Only the establishment of such a system will permit to develop and utilize planning models in the political field to an extent which will be indispensable in future.

10. Aspects for the establishment of supranationally combined social data systems in Europe

At present no other EC member state seems to dispose of such a comprehensive data bank system as the social information system of the Federal Government which is a genuine, continuously expanding instrument for the decision-making and planning processes. In the course of the development of the European Communities from an economic into a social union it will in the medium or long term prove necessary to set up corresponding or modified supranational information combination systems which permit a common action on a European level.

In the last few years this necessity was demonstrated in the following cases:

(1) Newspapers emphasized that the establishment of a "social computer" will be considered as the basic prerequisite for the development of a European social policy. "Extensive information on the social situation in the various countries is precondition for any central action. The Commission has already made the first step in this direction by announcing the establishment of a social computer. In future an electronic information system to which all member states are connected is to give rapid and reliable information on the labour market situation in the respective countries." (12)

(2) The Council of the European Communities decided to create a European foundation for the improvement of the living and working conditions. One of the primary objectives of this foundation is the establishment of an information and documentation system for the dissemination of relevant knowledge.

(3) The decision of the Council of the European Communities to establish a European centre for vocational training emphasizes that we should look for common features in this sector of a European employment policy. The exchange of information and experience will be one of the major tasks of this centre. Specific information is to be obtained on the basis of a selective documentation in the interest of a coordinated approach; the problems of vocational training structures are to be identified.

The social data bank of the Federal Minister of Labour and Social Affairs contains manifold data on all these fields.

There is a need for a number of clarifications in the national as well as in the international field in the interest of a harmonious development of a common information basis. The social information system of the Federal Republic of Germany has been conceived in such a way that some day it can be institutionally integrated into an overall information system of the Federal Government (Federal Data Bank). The more such national activities, coordination, data collection, and integration are advanced the more difficult it will be in the long run to bring about a functional coordination with international information systems for example with the different social information systems of the EC member states.

If such integration efforts are to be pursued on a European level it is indispensable to attain a certain degree of coordination in the development of a supranational social information system. The individual countries elaborate more and more binding identification and key systems and definitions. Any compatibility and standardization efforts which in the long run would bring about a standardization at a European level for a better mutual assessment of the state of social security, of the impact on employment policy at a European level or of socio-political structural measures, are doomed to failure without an integrated European basis for information and decision-making. Even at a national level the solution of tasks in the framework of social policy – as outlined above – are no longer possible without electronic data processing. The OECD untertakes efforts at an international level towards the development of generally recognized social indicators. All this demonstrates that in the supranational framework of the European Communities special efforts should be made for the clarification of these problems in order to guarantee the establishment of computerized information systems going beyond the conventional statistical methods and procedures.

The above-mentioned basic requirements were "highlighted" during the discussion in the German Parliament on the note of the community policy in the field of data processing (Parliamentary Document 7/1531), submitted to the Council by the EC Commission. The Parliamentary Committee for Labour and Social Affairs as well as the Economic Committee pointed out that the problems of electronic data processing in the field of social security had up to now played an only minor role in the activities of the Commission. "In view of the variety of the ever increasing data volume accumulating in the field of manpower and social policy in particular and in view of the growing need for such data for political decisions it is necessary to dispose also of reliable information systems on employed persons and social services in order to provide a basis for national and supranational manpower and social policies." Accordingly the Federal Government was requested to see to it that in its further negotiations with the EC Commission the following aspects be considered in pursuance of a European policy in the field of data processing:

1. European manpower and social policy must also be considered in the framework of the envisaged establishment of supranational information combination systems. To this end an early coordination of approach of all EC member states is indispensable.
2. The possibilities and the increased use of electronic data processing as well as its development are of primary importance for job security and humanization of working life. Therefore related tasks and problems should be included into the further considerations of the Commission on a community policy in the field of data processing.
3. In view of the manifold tasks resulting from the use of electronic data processing in the field of social security in all member states it is appropriate to set up a body dealing with 'EDP in manpower and social policy' in analogy to the body suggested in number 41 of the above-mentioned note of the Commission; this body shall have the tasks to
   - structure common problems,
   - propose development projects with an international character in the field of EDP application,
   - discuss the problems of classification and standards for EDP application and to elaborate proposals for their solution,
   - submit a documentation for a compatible data collection with the aim of establishing in future a supranational data combination system,
   - compile material for the report on the activities of electronic data processing in the Community envisaged by the Commission (No. 45 of the note).

The suggestions put forward by the German Parliament have meanwhile been submitted to the EC-Commission. Now it remains to be seen in which form the EC-Commission will take account of the proposals outlined above.

11. The social data bank – an information monopoly?

The multitude of possibilities offered by the data volume stored at the social data bank will doubtless contribute to an improvement of transparancy in our complex social and working life. In the interest of a quality improvement of political decisions, of the services rendered by the state to its citizens, of the documentation which is important for scientific and technical purposes, the social data bank will, in the years to come, be gradually extended to a social information system. This will permit to benefit from all advantages resulting from the utilization of electronics and valuable for the development of society. This means that in future, besides the conventional evaluation of statistics and reports, also the supply of specific evaluations and information will become possible by means of an improved communication with the social data bank. The increasing usability of the EDP-system will make a decisive contribution towards this end.

The Federal Minister of Labour and Social Affairs is determined not to create an information monopoly and withhold the information possibilities of the social data bank from the general public. It is rather intended to derive from this data volume all possible information in an anonymized, aggregated, regionalized, and technically classified form and to make it available to the general public, having due regard to the requirements of data protection.

The manifold chances and possibilities offered by such a social information system based on the data volume provided by the social data bank will be made use of by further developing it into an independent technical system, which, in a first stage, will contain prevalently numerical data on "manpower and social security". The development of such a social information system at the Federal Ministry of Labour has, first of all, the objective to improve the quality of life, to humanize work, to increase the benefit from governmental measures in the interest of society and to broaden and intensify the information of the community at large with a view to achieving more justice and equality of opportunity.

Literature

(1) Federal Minister of Labour and Social Affairs (Ed.): Social budget 1974. Bonn (1974), 33.
(2) Winkler, Paul: Die Sozialdatenbank. In: Computer und Gesellschaft. Published by F. Krückeberg, W. Walder, and B.A. Brandt. Stuttgart, Frankfurt 1974, p. 105.
(3) Zschau, Herbert: Sozialdatenbank – Ein Instrument der Sozialplanung. In: Data Report, 1974, No. 3, p. 4.
(4) Hartleb, Uwe: GOLEM informiert. In: Bundesarbeitsblatt, 1975, No. 2, p. 99.
(5) Geißler, Ulrich: Modellrechnungen zur Vorbereitung sozialpolitischer Entscheidungen. In: Datenverarbeitung in der politischen Planung. Published by Peter Hoschka and Uwe Kalbhen, Frankfurt/M. 1975.
(6) Katzer, H.: Aspects of modern social policy. Stuttgart, Berlin, Köln, Main (1969) p. 97.
(7) Berié, H.: The social Budget, foundation, methods, and procedures. Bonn-Bad Godesberg (1970).
(8) Ulrich, Erhard: Methods and approach of manpower demand forecasts and personnel planning. In: Handbuch der Personalplanung, edited by Herbert Schmidt, Hasso Hagenbruck, and Werner Sämann, Frankfurt/M., New York, 1975, p. 230.
(9) Eisner, M.: A Researcher's Overview of the TROLL/1-System. Cambridge (Mass.), 1972.
(10) Hoschka, Peter and Uwe Kalbhen: DV-Einsatz in der politischen Planung. Frankfurt, 1975.
(11) Grochla, E. and N. Szyperski: Modell- und Computergeschützte Unternehmensplanung. Workshop IV, Software für die Unternehmersplanung, Wiesbaden 1973.
(12) "DIE ZEIT", No. 21, 18 May 1973, p. 7.

Hospital Information System Hanover

Werner Weingarten, Hanover

I. General introduction

The integrated computer system of the hospital of the Medical School Hanover is called the MSH (*M*edical *S*ystem *H*anover). It is developed and operated by the Department of Biometrics and Medical Information Sciences of the Medical School Hanover. The Medical School is an independent college of medicine and not part of a university. It was founded in 1964 and the first students were admitted in 1965. The first inpatients were admitted in July 1971. Now, there are about 800 beds and the number of admitted patients varies from 70 to 100 per day. The overall projected total capacity of the Medical School Hanover is approximately 2000 beds. The rate of all patients is about 250 per day at *the moment*. The city population of Hanover provides the majority of all admitted patients. The Medical School is owned and operated by the State of Lower Saxony.

II. Definition of the general goals of the hospital information system in the Medical School Hanover

The Medical System Hanover has to carry out both hospital and patient management. The main functions of the integrating computer systems are:

1. communication
2. integration
3. system flow control.

It is very difficult to define a hospital in general terms. As an institution of academic learning, various projects of computer aided instruction, computer aided testing, procedural training, and scientific analyses as well as special research projects have to be supported. The tasks of a hospital computing system fall in the overall areas of information science which are:

1. data acquisition
2. information validation
3. information management
4. information evaluation
5. information (or system) flow control.

It is intended that the Medical System Hanover should serve the following functions:

1. data collection
2. communication
3. documentation
4. information integration
5. information presentation
6. information derivation
7. quality control
8. optimization of functions
9. general administrative functions.

The problems of medical information science and the multiple tasks originating from a large hospital dictate the objectives of the Medical System Hanover:

1. to provide a system for information acqusition, integration, retrieval and analysis concerning patients and human and material resources of the Medical School Hanover teaching hospital to support patients, hospital management and scientific research,
2. to support teaching activities and the academic administration of the Medical School Hanover,
3. to provide services and systems for scientific data analyses,
4. to provide the basis for research and teaching in the area of medical information science.

In this list, the different tasks are not necessarily compatible in their priorities. The system concept has to cope with this problem.

III. The general system concepts

The basic vehicles to perform the functions are;

1. a teleprocessing system
2. an integrating data bank
3. a timesharing system that will furthermore permit the simultaneous operations of different computer systems to perform the tasks in the various areas.

The hardware

In 1969, an IBM 360/67 was chosen. The following description is a short and not comprehensive one:

512 k Bytes of real core,
3 selector channels,
2 IBM 2301 drums with each 4 Mill. Bytes of storage,
3 disk units IBM 2314 with 8 accessable packs each (600 Mill. Bytes),

6 tape units,
8 IBM 2260 cathode-ray-tubes, 25 SIEMENS 8151, 22 IBM 2740 typewriters,
9 IBM 2741 devices as consoles for virtual machines,
19 IBM 3270 terminals,
1 optical reader and 1 paper tape reader,
1 offline plotter,
1 OCB unit (offline).

The software

The central supervisor is a system CP-67. The teleprocessing monitor is the Baylor Executive System for Teleprocessing (B.E.S.T.), which has been completed by a special set of PL 1 macros to serve the application programming. To build and maintain the integrating data base, the IBM Information Management System (IMS) has been chosen. This data base system specially permits the design of hierarchical data bank structures.

The general data base concept

The data base concept distinguishes between two types of data:

1. permanent (summary)
2. temporary (details).

'Permanent' means remaining on random access devices. The daily input to the data base system is kept in intermediate files for speed and data security reasons. The update into the IMS data base system is done in the evening, while in the meantime a log is kept where the information can be found in order to provide transparency for the user. The integrating data item of all patient related information is the identification number. The identification number has the structure: DDMMYYNNSX, where DDMMYY stands for day, month and year of birth and NN is a two digit number code based on the first two letters of the birth name. This code has been designed to achieve an equal distribution. S stands for the sex of the patient. In those cases (approximately 1 per 6,000) where two different patients coincide in the first 9 digits of their identification number, X will be incremented in order to differentiate them.

Separation of hospital functions

The main hospital functions are hospital management (administrative services), patient management (patient oriented services), and medicine (medical tasks and medical science, public health). Hospital management provides general administration and allocation of resources and facilities. Patient management deals with the individual patient before, during and after his stay in the hospital. Medical tasks and medical science are the specific medical activities in dealing with the patient as an individual and in general. The word management in this sense means:

1. coping with the increasing quantities of information
2. optimizing functions
3. the decision making, and
4. justifying of cost by effectiveness.

The patient's medical record

The principle governing storage is to keep condensed relevant information online as long as possible.

The content summary contains for example the diagnoses, risk facts, administrative and personal data.

Various statistical evaluations have been performed from the content summary, some of which are

- age distributions of patients in different clinics
- number of treatments per patient
- coding of risk facts
- diagnosis versus time of the year (season)
- hospital stay per diagnoses
- spectrum of coded diagnoses
- age distribution for different diagnoses.

The data bank part for actual patient data contains for example episode oriented information about the ward, the doctor, reportings and laboratory data. It is intended to keep these data for one treatment period. Data then are unloaded on tape in specific IMS program structures so that they may be reloaded if necessary. The actual patient data file was for example used to produce weekly accumulated laboratory reports whenever the patient got new laboratory data. This file enables the user to get quick on-line information about a patient. The accumulated laboratory report is available by visuable display units.

The data base system hat not replaced the manual chart, which is microfilmed and destroyed after the patient's discharge. The data base system does not contain enough data to be comprehensive.

IV. Administrative functions

1. *Admission*

The central and on-line admission of all inpatients consists of up to 72 questions dealing with the patient's address, insurance type, employer, secondary carrier, relatives, referring physicians, referring hospital and so on. After the dialogue, an aluminium foil is embossed on a typewriter terminal containing the most basic administrative information and the patient's identification number in a computer readable binary coded form. Thus the information can be transcribed on all kind of documents including optical reader forms. At the same time, stickers for various purposes are printed with the patient's name, ward number, and identification number. A request for insurance coverage is printed in a standard way and made ready for signature for the physician on the ward. Since the opening of the hospital, every inpatient has been registered in the MSH on-line.

2. *Billing*

The number of days spent in the hospital is the basis for billing, so that the admission and discharge procedures are able to provide the necessary information for the billing.

3. *Personnel*

Information about all personnel of the Medical School is recorded for supporting personnel management as well as the activities for the elections for the self-governing bodies of the Medical School.

4. *Other services*

Telephone service administration and accounting is done. An on-line program is available for budgetary purposes. Naturally, all these systems are data protected.

V. Service units

1. *Clinical laboratory*

The clinical laboratories are independent of the Department for Biometrics and Medical Information Sciences. Laboratory examinations are requested on optical reader forms. The link between the laboratory computer and the central computer is established to deliver identification information to the satellite. Via tape, the laboratory results are fed into the central computer.

2. *Pathology*

An on-line system has been developed for pathology reportings. Cytology and tissue pathology are done completely on-line. A thesaurus has been developed cooperatively between major European institutes for pathology in German speaking countries. All reports may be reviewed on-line and a hardcopy is produced and transmitted to the sender.

3. *Radiology*

Radiology modules, which have been developed together with Lodwick and Templeton have not yet been used within the MSH.

4. *Pharmacy*

The pharmacy system of the MSH is still under development. It accounts the drug movements to and from the wards. The system now provides the total accounting of the pharmacy. The drug inquiry system is developed, thus providing information about action and side effects, contra-indication, indication and incompatibility. Thus, in a later period the drug profile and the patient can be matched against these data in order to optimize therapeutic actions.

5. *Physiology*

The routine analysis of electrocardiograms is implemented by the group Schneider/Zywietz. They modified programs of Pipberger. A research project deals with the automatic analysis of electroencephalograms (Kuenkel). Both applications are not in full routine use. ECG reporting is done by the FTSS (Free Text Synthesis System) and es adapted to the new AMAP (*A*llgemeines *M*arkierungsbogen-*A*uswertungs*p*rogramm) system of the central computer.

6. *Patient dietary*

Patient menus are prepared on an automated tray-line. The employees put the different ingredients on the trays according to light-singals in front of them. The necessary control cards are produced by the MSH on a daily basis. Updating of diet plans is done by optical reader forms or on-line.

VI. Clinical units

1. *Nurses station system*

The dietary system is operated from the ward. Inquiries can be made from the wards and patients may be admitted. Research is being done in this area.

2. *Patient scheduling and hospital admitting*

There is no automated scheduling of patients now. Planning data may be extracted out of the admission data. A special evaluation for example shows the number of admissions per day, the number of discharges per day and the total number of beds occupied. The total number of bed occupancy shows remarkable dips during the time of Christmas, Easter, Pentecost and 1st of May (general holiday). Furthermore, a decrease of beds occupied could be seen this year during the beginning of the general vacation in the state. It is remarkable that after the decrease no compensation is registered. There are weekly ondulations too, peaks at the beginning of the week overload human and material resources during the first part of the week or lead to poor utilization at the end of the week.

3. *Medical information processing*

3.1 Automated patient history

CDSS (Clinical Decision Support System) has been adapted to the Medical System Hanover. The relevant questionnaire contains of approximately 400 questions.

3.2 Physical examinations

The FTSS (Free Text Synthesis System) uses optical reader forms which may be used as primary documentation in the patient's chart, and which specify in a command language how the findings should be printed and synthesized either to a description of the physical findings and/or a consultation report.

3.3 Problem oriented records

The test application on pediatric wards allows for specification of problems, critical evaluations of problems, updating, and free text adding to reports to be generated from optical reader forms with the FTSS. Discharge summaries (problem epicrises) are generated.

4. *Special inpatient services*

Ancillary systems are those systems which support directly the physician's activity in dealing with the patient. In this category fall: special inquiry systems, decision support systems, interpretation aids (discipline oriented, problem oriented), problem management systems (fluid balance calculation, flow rates, heart and lung diagnostics, simulation systems, etc.) and aid in specific research projects (data acquisitions, grouping and analysis).

4.1 MHBS (*M*ineral*h*aushalts*b*ilanzierungs*s*ystem)

This system performs the balance calculations for intensive care patients, especially for those recovering from neurosurgery.

4.2 AKOS (*A*llgemeines *K*odierungs*s*ystem)

This system allows the encoding of diagnoses by a free text entry of five or more letters of a significant word in a diagnostic term. From a table lookup all matches are displayed. The user specifies the most significant match, which is added to the patient's record.

4.3 Diagnostic aid

There exists a minimum of research efforts, but not yet routine procedures in the Medical System Hanover. Two models are taken, a probabilistic one called CADS (*C*omputer *A*ided *D*iagnostic *S*ystem) and the deterministic model with the neuron-like concept of CDSS (*C*linical *D*ecision *S*upport *S*ystem). The basis of the probabilistic model is the Bayesean theorem. Models are defined for thyroid disease, bone tumor, gastric ulcer, heart disease and solitary lung nodules. A model for acute abdominal symptomatology is under development. A deterministic model is implemented for the interpretation of blood gas analyses. Laboratory values are entered, and then diagnostic recommendations are calculated, including explanations as to possible causes and recommendations for therapy. Another module has been implemented for the differential diagnoses of bone marrow smears.

4.4 IRSD (*I*ndividual *R*etrieval *S*ystem for *D*iagnoses)

Diagnoses are inverted for a smaller subpopulation to cover special areas and research projects. Especially categorical retrieval is established. The system uses the coding mechanisms of AKOS.

4.5 PIDS (*P*atient *I*nformation *D*isplay *S*ystem)

This system is the inquiry system to access the integrated data base. The system's output is patient oriented. There are two main tasks:

1. to identify a patient out of the total population of nearly 70,000 patients of the data bank and
2. to deliver information about an identified patient (options).

The identification of a patient is possible by using the following search criteria: identification number, emergency identification number, admission number, variable length substring of the name, birth date, date of admission, transfer or discharge (optically qualified by ward specification), and actual ward number. Generally, the user must choose the patient from a list. The immediately available options are: administration data, summary, medical risk data, diagnoses, laboratory data and a synopsis of the medical record. The access to different subsets of the complete information is controlled by a personal password which covers the spectrum of up to 64 different access methods. A log is kept of all users and their transactions.

4.6 There are two systems which are generally adaptable to user wishes and which perform the variable input of user data (clinics for example) in the data base.

4.6.1 AMAP (*A*llgemeines *M*arkierungsbogen-*A*uswertungs*p*rogramm)

This system is a general program for mark sense form evaluation which is adapted by compilation steps for various purposes. The system provides plausibility checks, statistical functions, and free text additions on-line to mark sense forms which cause the writing of letters for example. This system is in routine use with 13 different applications, in routine test with 6 applications and in the design phase for further 7 applications. The system's functions enable the replacement of the older FTSS system. The normal input is done via mark sense forms, but there exists an on-line version too.

4.6.2 DADIMOPS (*Da*ta *Di*rected *M*edical *O*riented *P*rocessing *S*ystem)

This system is adapted to the various problems by data which control the program flow. This system is an on-line system and in routine test with one application.

4.7 GORS (*G*eneral *O*n-line *R*etrieval *S*ystem)

New tools are developed to provide general data directed input/output systems for all kinds of data bases with general defined plausibilities. These systems use data base descriptions, a data definition language, and a kind of data manipulation language. The general on-line retrieval system is in a routine test, whereas the other modules are still under development.

5. *Outpatient clinics*

Outpatient clinics are treated in the same way as the main hospital. The patient-specific identification number allows for a comprehensive record linkage. At the moment, outpatients are entered into the main data base using off-line procedures. Routine tests are performed in the pediatric wards to admit outpatients by a shortened on-line dialogue.

6. *Physician consultation*

Here the FTSS system is used. There exist modules for gynecology in routine use, modules for ear, nose and throat consultation, traumatology and dentistry are under development.

VII. Education and Research

The Department for Biometrics and Medical Information Science provides courses for students and faculty members in various programming languages and techniques, and instructions in biomathematics and basic statistical procedures. Furthermore, there are specific tasks defined, as there are the teaching of the most basic principles of medical information science and data processing and the introduction to existing systems.

1. SIS (*S*equential *I*nstruction *S*ystem)
This system is a tool for computer aided instruction. Questions are given in the multiple choice fashion, branches in the dialogue depend upon correct or wrong answers. Instruction modules are implemented for medical contents as well as programming and system concepts.

2. Computer-aided testing
In seminars and examinations students may give their answers on optical reader forms so that an overall evaluation can be made. The program generates output for the instructor as well as for the students. The students may be grouped into four groups with permutated questions and the questions may have different weights. Difficulty, separation, and selectivity of questions are evaluated.

3. PIRS (*P*ersonal *I*nformation *R*etrieval *S*ystem)
This system handles the storage and retrieval of literature. There exist on-line and off-line modules which allow searches for character strings by the author, title or bibliography section of a literature citation, key word retrieval, and combination of both. The literature may also be grouped for reference lists in publications.

4. Statistical and mathematical systems
The off-line-library covers the most wellknown packages. For CMS (Cambridge Monitor System) a conversational system is available which allows formulation of algorithms and mathematical models.

5. The content of the different data bases has been statistically evaluated from numerous points of view, for example concerning laboratory analyses:
- frequencies of analyses in different clinics
- utilization of the spectrum of possible analyses
- number of analyses per treatment
- arithmetical means, standard deviations, broken down by the sex, the age, and the diagnosis of the patients
- simple correlations between laboratory analysis and coding diagnoses
- presentation of the results of a set of analysis belonging to the same diagnoses in a suitable recognizable two dimensional plot.

Contributions of Information Science to Data Documentation

Gernot Wersig, Berlin

1. The basic concept of "information"

The term "information" is used in a large variety of ways so that it is nearly impossible to understand what is meant by it if the term is used undefined in scientific communication. A semantic analysis shows that there are at least six different approaches to the use and understanding of the term in the whole field of disciplines. All uses may be justified but in a specific case of application it has to be specified which use is being made,

a) The *structure approach* states that already the structure of the world, which may or may not be perceived by man are "information". "Information" thus is independent whether a human being gathers it or not (used in philosphy).

b) The *knowledge approach* states that the knowledge built up on the basic of perception of the structure of the world is "information" (used e.g. in decision theory).

c) The *message approach* states that a message itself (without its meaning) is "information" (widely used in mathematical communication theory).

d) The *meaning approach* states that the "meaning" of a message (whatever this may be) is its "information" (used in electronic data processing. In linguistics sometimes the reverse happens).

e) The *effect approach* states that "information" is a specific effect a message exerts on the receiver (e.g. reduction of uncertainty).

f) The *process approach* states the whole process of processing and transmitting of data may be called "information".

2. The understanding of "information science"

The term "information" shall be understood as being part of the compound "information science" according to the effect approach, because this seems to be the appropriate one for the domain of study. "Documentation", "information retrieval", "information work" appeared at a stage when scientific and technical communication were disturbed by the growing number of printed items and the growing complexity of formal communication channels. The ultimate purpose of all information activities therefore is to reduce the disturbance of scientific and technical communication by directing data or sources of data to the receivers who need them and make optimal use of them.

"Information science" therefore becomes the discipline concerned with the organisation of communication processes in such a way that optimal information transfer will be effected. In the meantime the area in which information science works is quite large. The basic understanding stems from literature documentation but is valid for data documentation and information as well (documentation being the more passive, information being the more active part of the whole purpose of information and documentation agencies).

3. Different types of agencies

The new understanding shows that there are a lot of different types of agencies being concerned with information work. They all have to be seen as fulfilling different functions in the whole system of activities being connected by the general aim of fulfilling the information needs of users. Such functions are e.g.

- management of records: museums, archives, libraries
- information about records: special libraries, information/documentation agencies
- evaluated information about records: information analysis centers
- management of data: data archives, data centers
- evaluation of data: information analysis centers, information evaluation centers
- information about ongoing projects and recent expertise: project documentation
- integration of functions: hospital information system, management information systems, planning information systems etc.

4. Two basic types of information/documentation

a) direct documentation (data documentation): a question about something is answered directly
b) indirect documentation (literature documentation): a question is not answered directly but by references to sources where possibly an answer may be obtained.

Sometimes the difference between them is said to be the difference between "hard" and "soft" data. But this is simply not true: data documentation can be concerned with hard and soft data (e.g. blood pressure, specific talents of people, definitions) and literature documentation is obviously concerned with hard data as well. But there are some information needs better served by data documentation (questions which can be answered simply by a formatted response) or by literature documentation (questions concerning concepts, thoughts, ideas etc.).

This makes obvious that both approaches cannot be distinguished by reducing them to different sources: methods of literature documentation can be applied to non-published materials (e.g. patients records) as can methods of data documentation be applied to published literature.

5. Similarities and differences between literature documentation and data documentation

a) Both serve the same purpose: satisfaction of user needs. But both serve them on different levels (references and accompanying additional information vs. data).

Excursion: The role of the user

Sometimes it seems that information agencies are only working for their own sake. But we have to realise that the ultimate goal is fulfilment of user needs. The most widely used approaches to the phenomenon "user" are

- the user is looked at as being noise for the information agency
- the user is treated like a silly little child who has to be guided by the information specialist
- the information specialist compares the user's abilities with his own.

None of these approaches is appropriate. Users are people having specific occupational roles in their work. They try to get the information they need by the easiest way – and this is not the information agency but the direct contact with a colleague (about 50% of all information searches). The problem is that personal communication with other colleagues is efficient in some cases but not in all (e.g. if thorough searches in data bases are necessary, if the latest data are required). The information agency therefore should try to find out in which cases its clientele uses the direct contacts inappropriately, and then convince the user of the wisdom of using the services of the information agency instead.

Some basic rules for information specialists are:

- He should try to find out who his users are personally (and who should be a user).
- He should try to find out their information needs.
- He should try to know the personal contacts they usually use (e.g. he may look at cases where they failed to get the appropriate information and then demontrate how he could have solved the problem; he may design some pieces of the information service according to their needs and convince them that they really do serve their needs).

It becomes obvious that work in data documentation requires more responsibility towards the user than in literature documentation. In literature documentation the user receives some references out of which at least some may be relevant and therefore he will have some satisfaction. In data documentation often only one item of data is needed and if this is not correct or satisfying the agency looses its credibility. In data documentation, therefore, one has to look more thoroughly for the specific user needs and has to take even more care to accurate services.

b) The *documentary processes* used in data documentation and literature documentation basically are equivalent: data are the raw material, they have to be transformed into documentary data (that means data presumably effecting information for the clientele of the specific agency), and transmitted to the user in order to effect the required information.

But there are basic differences in some important stages.

c) The difference is obvious in acquisition of input into the information system.

Literature documentation traditionally accepts what has been produced but does not try to control literature production (although, in principles, this could be a purpose of literature documentation as well). In data documentation acquisition of data is best performed at the moment of data production (e.g. medical data). This means that data documentation is much closer to the production process than literature documentation and sometimes the performance of data documentation requires the alteration of data production structures.

If data are required as in literature documentation – in a secondary way, a very important difference to literature documentation occurs: Evaluation in this case is unavoidable, since the user trusts data documentation much more (and has to).

d) In literature documentation *documentary language* have to overcome the ambiguities of natural language. This means that there are artificial languages being designed to act as a kind of standard referece code. Descriptors are labels to equivalence classes of terms being designed to indicate a class of concepts.

In data documentation documentary languages are used to overcome the difficulties of different definitions of data categories (which may have been made explicit or not). Descriptors are used merely to act as labels for a set of characteristics (particularly with statistical data), which usually form an artificial concept. The artificiality in literature documentation consists in term representation, the artificiality in data documentation consists in concept construction (e.g. "XYZ but excluding u, v, w").

Excursion: Traditional documentary languages in data documentation

Traditionally the following approaches to documentary languages are known:

- natural language based:
  - keywords
  - subject headings
  - thesauri
- artificial language based:
  - classifications with notational systems
    - hierarchical
    - faceted
  - numbering systems

In data documentation natural language based documentary language are not as widely applicable as artificial language based systems. This becomes obvious when looking at the internal structure of data which usually consists of three main data elements:

- data category: explaining the contents of the data units,
- constituent variable: explaining the main value of the data unit,
- determinant variable(s): explaining additional values necessary to understand the value of the constituent variable appropriately.

The variables usually consist of

- value (e.g. numerical value)
- category (e.g. ° in connection with temperatures)
- dimension (e.g. C in connection with temperatures)

If we have a file consisting of several data categories (that means data of different kinds), the entry element into the file would be the data category. The more elements of the whole data unit (e.g. categories and dimensions of variables) are included in the description of the data category, the less redundant is the representation of the other elements (values of variables). This means that the description of the contents of data units by the data category unavoidably becomes rather complex.

If we collect data for statistical purposes (not single data units but assemblages of data units with the same charateristics are of interest here) a classificatory approach may be worth-wile: large and concise class descriptions (data categories), associated codes (or notations) for brief representation, hierarchies to aggregate data on higher levels.

A regrouping of data units, according to subsets of data categories or other elements is impossible to achieve when using the above method. This regrouping will be permitted if another classificatory approach is used, the faceted one: data files will not be built up for the data unit, as a whole, entering it only by its complex data category, but by a set of related inverted files where each variable (or even elements of the variables) will enter individually (internally linked). Now regrouping in terms of e.g. constituent and determinant variables is possible (in terms of medical diagnoses: if diagnoses are classified by facets like localisation, pathic process, etiology then regrouping according to the specific facets is possible).

The hierarchical approach allows allocation to statistical classes; the faceted approach allows secondary evaluation.

In both approaches item retrieval is difficult; several items are assembled in one class, the selection of single items then is a sequential process. An item oriented data documentation system is better organized by a detailed description of the contents components of the respective data units. This is best done by a thesaurus approach allowing two functions:
- if used as the basic description medium it allows item retrieval
- if used as a description medium where classes of classifications are allocated to a specific set of descriptors, statistical evaluation according to flexible hierarchies is possible.

This shows that in data documentation – where usually several functions have to be performed on the same data basis – it is more necessary than in literature documentation to combine all three approaches.

e) *Information services*. Data documentation and literature documentation both could use the same structure of information services: SDI, compilations, retrospective retrieval, continuous information publications.

But data documentation services must be much more synthetic. In classical literature documentation services the items are singular, synthetization is only introduced by inverted files (indexes). In data documentation single data items do not say anything. They have to be aggregated, synthesized. Furthermore: In literature documentation information systems, evaluation of specific programmes, new combinations of findings etc. (if any) is usually done by the brain of the user. In data documentation systems we have to take care that we provide the user with possibilities to connect retrieval of data with flexible possibilities of regrouping, aggregation, evaluation, simulation models. This cannot be done solely by the brain, the system has to provide assistance also.

6. Integrative needs

Users' needs do not care for organizational or ideological differences between information workers; they require integrative systems. In future we have to foresee information agencies in which data documentation and literature documentation have to be combined in integrated systems. This is a challenge much more for data documentation than for literature documentation: In spite of national, international or interdisciplinary cooperation in literature documentation and data documentation, literature documentation activities are more used to meeting conventions. Thus data documentation activities have to enter the stages of cooperation more than they do at present.

Only close connection between literature documentation and data documentation will allow the information systems of the future (e.g. planning information systems) and a closer relation of information activities and research to the development as a whole (since data documentation has a much closer relation to this area than literature documentation). Data documentation is at the same time an integral part of scientific/technological work and of information/documentation work.

Bibliography

1. General Views, Compendia, Directories
2. Basic Software Design of Data Bases
3. Acquisition of Data Including Quality Control
4. Management and Organizational Aspects
5. Confidentiality and Security of Data
6. Information Retrieval
7. Information Analysis, Evaluation of Data
8. Presentation of existing Data Banks
    - Administration
    - Law and Police
    - Social
    - Traffic
    - Chemistry
    - Engineering
    - Geo-Sciences
    - Medicine and Biomedicine
    - Standards
    - Miscellaneous

1. General Views, Compendia, Directories.

1.1 AEC and DOD information analysis centers. Spec. Libr. 57 (1966), 21–34.

1.2 Cole, R.I.: Data: information availability. Washington, Thompson Book Company (1966).

1.3 Committee on Data for Science and Technology of the International Council of Scientific Unions (CODATA): International compendium of numerical data projects. A survey and analysis. Berlin – Heidelberg – New York, Springer (1969).

1.4 Date, C.J.: An introduction to data base system. Addison-Wesley (1975).

1.5 Gesamtverzeichnis zum Thema Faktendokumentation, Datenzentren, Datenbank. Rechentechnik, Datenverarbeitung 5 (1969). Beiheft 2, 57–63.

1.6 Institut für Datenverarbeitung; Fachbereich Information: Faktendokumentation, Datenzentren, Datenbank. Dresden, Institut für Datenverarbeitung (1969).

1.7 Institut für Rationalisierung und Organisation der Elektroindustrie (Eds.): Faktendokumentation und Datenbänke. Berichtszeit: 1964–1969. Dresden, Inst. f. Rationalisierung u. Organisation der Elektroindustrie (1969).

1.8 Koller, S.L.; Wagner, G. (Eds.): Handbuch der medizinischen Dokumentation und Datenverarbeitung. Stuttgart, Schattauer (1975).

1.9 Office of Critical Tables: Directory of continuing numerical data projects. A sruvey and analysis. Washington, National Academy of Science – National Research Council (1961).

1.10 Price, G.F.: Les dix commandements de la base de données. L'informatique (Paris), 47 (1974), 10–21.

1.11 Science-Communication: Plan for study and implementation of national data system concepts (Study of scientific and technical data activities in the United States, Vol. 1). Springfield (Va.), CFSTI (1968).

1.12 Simpson, G.S. Jr.; Murdock, J.W.: Qualitative approach to scientific information problems. Columbus (Ohio), Batelle Memorial Inst. (1960).

1.13 Taraboi, V.: Concerns and achievements in the area of spreading information regarding industrial proceedings and products. Report presented at the Seminar of the UN Economic Committee for Europe Bukarest. Probl. de informare si documentare, Bukarest 12 (1972), 5–13.

1.14 Wooster, H.: The future of scientific data. OAR Research Review 7 (1968), 15–17.

2. *Basic Software Design of Data Bases*

2.1 Abrahamsson, S.; Larsson, K.: Danderyd hospital computer system. 3. basic software design. Comput. biomed. Res. 4 (1971), 126–140.

2.2 Babatz, R.; Bungers, D.; Rüdiger, N.: AKAM – ein Dialogprogramm zur Erstellung, Bearbeitung und Analyse von Bevölkerungs- und Erwerbstätigenprognosen am Bildschirmterminal. GMD-Spiegel (St. Augustin) 4 (1975), 21–27.

2.3 Barnett, G.O.; Greenes, R.A.: Interface aspects of a hospital information system. Am. N.Y. Acad. Sci. 161 (1969), 756–768.

2.4 Barnett, G.O.; Greenes, R.A.; Grossman, J.H.: Computer processing of medical text information. Meth. Inform. Med. 8 (1969), 177–182.

2.5 Baum, C.; Gorsuch, L. (Eds.): Proceedings of the second symposium on computer-centered data base systems. Santa Monica (Calif.), System Development Corporation (1965).

2.6 Berztiss, A.T.: Data structures. Academic Press (1973).

2.7 Cagan, C.: Data management system. Melville (1973).

2.8 CODASYL Systems Committee: Feature analysis of generalized data base management systems. (1971).

2.9 CODASYL Data Base Task Group: April 71 report. (1971).

2.10 Codd, E.F.: A relational model for data for large shared data bases. CACM 13 (1970), 5.

2.11 Dathe, G.; Dreckmann, K.-H.; Henselmann, G.: Designs principles of a data base system for unlimited quantities of data with highly complex structures. International Conference on Very Large Data Bases, Boston (1975).

2.12 Dirlewanger, W.: Programmiersprachen bei Stapelverarbeitung bzw. Dialogbetrieb. In: Dirlewanger, W.; Dobler, K.-U.; Hieber, L.; Roos, P.; Rzehak, H.; Schneider, H.-J.; Unger, C. (Eds.): Einführung in Teilgebiete der Informatik I. Berlin, New York, E. de Gruyter (1972), 70–80 = Sammlung Goeschen, Bd. 5011.

2.13 Dreckmann, K.-H.; Hofmann, G.: Ein Programmsystem zur Erfassung von Daten aus komplex strukturierten Tabellen. Gesellschaft f. Informatik, 5th annual meeting (1975).

2.14 Engles, R.W.: A tutorial on data base organization. Automatic Programming, Vol. 7, Part 1 (1972).

2.15 Gibbons, G.D.: Data structures for question answering systems. San Diego (Calif.), Naval Postgraduate School (1972).

2.16 Hartleb, U.: GOLEM informiert. Bundesarbeitsblatt 2 (1975), 99.

2.17 Hollederer, K.: SESAM Verfahrensübersicht. München (1974) = Siemens-Schriftenreihe data praxis 150.

2.18 IBM: Information management system/360 for the IBM system/360 (system description) version 2. General Information Manual.

2.19 STAIRS. Storage and information retrieval system. Ein IBM Dokumentationssystem. Bonn, IBM, Form M12-1016 (1971).

2.20 Lutz, Th.: Die Datenbank im Informationssystem. Oldenburg (1971).

2.21 Lutz, Th.: Das Computerorientierte Informationssystem. De Gruyter (1973).

2.22 Lutz, Th.: Die Mathematik der Datenbank. IBM Nachrichten (1975), 225.

2.23 Lutz, Th.: Informationssysteme und Datenbanken. IBM Nachrichten (1973/74), 215–218.

2.24 Merten, H.: Datenbank- und Dokumentationssysteme – ein Vergleich. Online (Köln) 12 (1974), 744–746.

2.25 Merten, H.: Datenbanksysteme und Dokumentationssysteme – ein Vergleich. Online (Köln) 12 (1974), 642–646.

2.26 Messina, C.G.; Hilsenrath, J.: EDPAC: utility programs for computer-assisted editing, copyproduction, and data retrieval. Washington, National Bureau of Standards (1969) = NBS Techn. Note 470.

2.27 Mittman, B.; Chalice, R.; Dillaman, D.: Mixed data structures in a multipurpose retrieval system. J. of the ASIS 24 (1973), 135- 141.

2.28 Nador, P.: Data administration and documentation associated with data base management systems. In: Samuelson, K. et al: Systems, cybernetics and information networks. Stockholm (1972), 34–52.

2.29 Rustin, R. (Ed.): Data base systems. Prentice Hall (1972).

2.30 Schnell, P.M.: ADABAS – Adaptable data base system-introduction. Darmstadt, Software AG (1972).

2.31 Senko, M.E.: Data structures and accessing in data base systems. IBM Systems Journal 12 (1973).

2.32 Skronn, H.J.: Methoden der Strukturierung von Datenbanken. Angew. Informatik 15 (1973), 204–210.

3. Acquisition of Data Including Quality Control

3.1 Adam, W.E.; Lorenz, W.J.: Erfassung und Bearbeitung der Daten in der Nuklearmedizin. In: Ehlers, C.T.; Hollberg, N.; Proppe, A. (Eds.): Computer: Werkzeug der Medizin. Berlin–Heidelberg–New York, Springer (1970, 1–14.

3.2 Boehm, K.; Thome, R.: Kosten verschiedener Methoden der Datenerfassung für die Literaturdokumentation. Nachr. Dok. 26 (1975), 68–71.

3.3 Bonnet, P.D.: The medical record – what is it? Bull. of the Amer. Soc. for Inform. Sci. 1 (1974), 8–35.

3.4 Brolin, I.: MEDELA – An electronic data-processing system for radiological reporting. Radiology 103 (1972), 249–255.

3.5 Burmeister, M.; Seyferth, A.; Teichmann, H.-J.; Eckert, R.; Encarnacao, J.: Die Formularsprache FOSPRA und ihre Implementierung im REDAS-Datenerfassungssystem. Angew. Informat. 6 (1975), 247–256.

3.6 Carter, N.W.; Griffiths, D.; White, C.J.: Computers in the laboratory: A problem of communication. Int. J. bio-med. Comput. 3 (1972), 147–153.

3.7 Engelbrecht, R.; Schertlein, G.; Reichertz, P.L.: AKOS – Allgemeines Kodierungssystem. In: Fuchs, G.; Wagner, G. (Eds.): Krankenhaus-Informationssysteme. Stuttgart–New York, Schattauer (1972), 95–100.

3.8 Engle, R.L.: The medical record – where is it heading? Bull. of the Amer. Soc. for Inform. Sci. 1 (1974), 9–11.

3.9 Geißler, U.: Modellrechnungen zur Vorbereitung sozialpolitischer Entscheidungen. In: Hoschka, P.; Kalbhen, U. (Eds.): Datenverarbeitung in der politischen Planung. Frankfurt (1975).

3.10 Giere, W.: Praxisrationalisierung und Dokumentation durch programmierte Befundschreibung. EDV Med. Biol. 4 (1973), 38–41.

3.11 Gösmann, D.; Scheibe, R.: Die Erfassung von Informationsnachweisen für AIDOS. Informatik 20 (1973), 24–29.

3.12 Grochla, E.; Szyperski, N.: Modell- und Computergeschützte Unternehmensplanung. Workshop IV, Software für die Unternehmerplanung, Wiesbaden (1973).

3.13 Hackl, J.; Mueller, K.; Schipper, P.; Sporn, C.: Maschinelle Durchführung der Qualitätskontrolle im klinisch-chemischen Laboratorium. In: Reichertz, P.L.; Holthoff, G. (Eds.): Methoden der Informatik in der Medizin. Berlin–Heidelberg–New York, Springer (1975), 145–158.

3.14 Heu, R.; Otten, J.; Purps, H.-D.: Ein autarkes Datenverarbeitungs-Subsystem für das klinisch-chemische Laboratorium. Röntgenstrahlen (1974), 25–31.

3.15 Hoehne, K.H.; Dahlmann, K.; Dix, W.-R.; Harm, K.: Computer application in the clinical laboratory – The information system approach with the system LABMAT. Meth. Inform. Med. 13 (1974), 226–232.

3.16 Hoschka, P.; Kalbhen, U.: DV-Einsatz in der politischen Planung. Frankfurt (1975).

3.17 Kanzler, G.; Giere, W.; Arndt, J.F.: Programmierte Befundschreibung und Auswertung proktologischer Untersuchungen über EDV. Z. Gastroent. 10 (1972), 567–574.

3.18 Koeppe, P.: ORVID. Ein Verfahren zur Automatisierung der Befunderhebung. Datenverarbeitung und Medizin. IBM-Seminar, Bad Liebenzell 1970, Bonn, IBM Form K 12-1012-0 (1971), 141–150.

3.19 Krieg, A.F.; Henry, J.B.; Stratakos, S.M.: Analysis of clinical pathology data by means of a user-oriented on-line data system. In: Enslein, K. (Ed.): Data acquisition and processing in biology and medicine. Oxford, Pergamon Press (1968), 163–172.

3.20 Laeser, S.: Programmierte Befundung mit RADIAR. Radiologe 14 (1974), 335–338.

3.21 Lippmann, E.O.; Preece, J.F.: A pilot on-line data system for general practitioners. Comput. biomed. Res. 4 (1971), 390–406.

3.22 Novak, D.: ROETEXT – ein System für die automatische Textverarbeitung in der Röntgendiagnostik. Radiologe 14 (1974), 277–285.

3.23 Pau, L.F.: Sequencing laboratory tests for improved sequential diagnosis. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDINFO 74, Preprints, 569 573.

3.24 Pendergrass, H.P.; Greenes, R.A.; Barnett, G.O.; Poitras, J.W.; Pappalardo, A.N.; Marble, C.W.: An on-line computer facility for systematized input of radiology reports. Radiology 92 (1969), 709–713.

3.25 Petsch, H.; Killian, K.P.; Knedel, M.: Zentrale Überwachung und Steuerung eines Laborsystems. In: Reichertz, P.L.; Holthoff, G. (Eds.): Methoden der Informatik in der Medizin. Berlin–Heidelberg–New York, Springer (1975), 75–81.

3.26 Thompson, G.E.: On-line data bank for admissions, laboratories and clinics. In: Bekey, A.G.; Schwartz, M.D. (Eds.): Hospital information systems. New York, M. Dekker Inc. (1972), 111–130.

3.27 Ulrich, E.: Methods and approach of manpower demand forecasts and personnel planing. In: Schmidt, H.; Hagenbruck, H.; Sämann, W. (Eds.): Handbuch der Personalplanung. Frankfurt/M., Herder u. Herder (1975), 230.

3.28 Wilson, J.; Brown, L.; Risso, W.; Knight, J.; Schultz, A.; Brown, F.: An on-line multi-instrument laboratory computer system. Int. J. biomed. Comput. 4 (1973), 135–148.

4. Management and Organizational Aspects

4.1 DATUM; Siemens; Stadt Köln: Datenverarbeitung für die kommunale Planung. Leopoldshafen: Zent.stelle f. Atomkernenerg.-Dokum. (ZAED), (1974) = Bundesministerium für Forschung und Technologie. Forschungsbericht FB DV 74-03.

4.2 Gekeler, O.; Herdt, K.D.; Oberender, W.; Ausschuß für wirtschaftliche Verwaltung in Wirtschaft und öffentlicher Hand (AWV): Warenkatalogisierung und Kommunikation über die Ware. Pullach, Verl. Dokum. (1974).

4.3 Haendler, H.: Bemühungen zur Schaffung besserer Unterlagen für Futterwertbestimmung. Wirtsch.eigene Futter 12 (1966), 16–27.

4.4 Pflug, C.; Kruse, P.; Schlegel, R.: SOPS-BASTEI beim Aufbau einer Methodenbank zur Unterstützung eines Leitungsinformationssystems. Rechentech. Datenverarb. 12 (1975), 57–61.

4.5 Seibert, G.: Datendokumentation. Methoden – Fakten – Entscheidungshilfen im Unternehmen. In: Deutsche Gesellschaft für Dokumentation: 2. Öffentliche Veranstaltung bei dem Deutschen Dokumentartag 1971. Herrenalb (1971), 23–29.

4.6 Stegemann, H.: Kostenrechnung einer Datenbank für Umfragen. Nachr. für Dok. 23 (1972), 69–79.

4.7 Sweeney, U.J.: Restructuring the NASA data base for time/cost savings. Spec. libr. (New York) 65 (1974), 499–504.

4.8 Vigler, G.J.: Beurteilungskriterien für Datenbanksysteme. BTA 12 (1971), 290–295.

5. *Confidentiality and Security of Data*

5.1 Anderson, J.P.: Information security in a multi-user computer environment. In: Rubinoff, M. (Ed.): Advances in computers, vol. 12. New York–London, Academic Press (1972), 1–36.

5.2 Aronson, E.: The social animal. San Francisco, Freeman (1972).

5.3 Asch, S.E.: Studies of independence and conformity: I. A minority of one against a unanimous majority. Psychol. Monographs Gen. Applied 70 (1956), 9.

5.4 Asch, S.E.: Opinions and social pressure. Scientif. Amer. 193 (1955) 5, 31–35.

5.5 Assakul, K.; Proctor, C.H.: Testing independence in two-way contingency tables with data subject to misclassification. Psychometrika 32 (1967), 67–76.

5.6 Astin, A.W.; Bpruch, R.F.: A "link" system for assuring confidentiality of research data in longitudinal studies. Amer. Educat. Res. J. 7 (1970), 615–624.

5.7 Babcock, J.D.: A brief description of privacy measures in the RUSH time-sharing system. In: AFIPS conference proceedings, 1967 spring joint computer conference vol. 30. Washington, Thompson Books (1967), 301–302.

5.8 Baran, P.: Communications, computers and people. In: AFIPS conference proceedings, 1965 fall joint computer conference, vol. 27. Washington, Thompson Books (1967), 45–49.

5.9 Baran, P.: On distributed communications: IX. Security, secrecy, and tamper-free considerations. Memorandum RM-3765-PR, The Rand Corporation, Santa Monica (Calif.), (1964).

5.10 Baruch, J.J.: Hospital automation via computer time-sharing. In: Stacy, R.W.; Waxman, B.D. (Eds.): Computers in biomedical research, vol. 2. New York–London, Academic Press (1965), 291–312.

5.11 Boehm, K.: Protection and confidentiality of medical data. II. Simple methods for meeting the users' needs. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDINFO 74, Preprints, 193–196.

5.12 Boehm, K.; Hahne, W.; Koehler, C.O.; Wagner, G.: Datenschutz durch rechnerinterne Vercodung. Z. Datenverarb. 10 (1972), 377–379.

5.13 Boruch, R.F.: Security of information processing – implications from social research. AFIPS Conf. Proc. 41, I (1972), 425–533.

5.14 Boruch, R.F.: Strategies for eliciting and merging confidential social research data. Policy Sci. 3 (1972), 275–297.

5.15 Boruch, R.F.; Endruweit, G.: Mathematische Methoden zur Sicherung der Vertraulichkeit und Anonymität von Forschungsdaten. Zschr. Soziol. 2 (1973), 227–238.

5.16 Browne, P.S.: Taxonomy of security and integrity. In: Hoffman, L.J. (Ed.): Security and privacy in computer systems. Los Angeles, Melville Publ. Co. (1973), 369–378.

5.17 Caroll, J.M.; McEllan, P.M.: Fast "infinite-key" privacy transformation for resource-sharing systems. In: Hoffman, L.J. (Ed.): Security and privacy in computer systems. Los Angeles, Melville Publ. Co. (1973), 139–156.

5.18 Conway, R.W.; Maxwell, W.L.; Morgan, H.L.: On the implementation of security measures in information systems. In: Hoffman, L.J. (Ed.): Security and privacy in computer systems. Los Angeles, Melville Publ. Co. (1973), 244–265.

5.19 Dinklo, J.A.: Confidentiality of medical data in the usage of data banks. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDINFO 74, Preprints, 181–187.

5.20 Dowling, T.A.; Shachtman, R.H.: On the relative efficiency of randomized response models. J. Amer. Statist. Ass. 70 (1975), 84–87.

5.21 Eimeren, W.v.; Selbmann, H.-K.; Überla, K.: Modell einer allgemeinen Vorsorgeuntersuchung im Jahr 1969/70. Schlußbericht. Bonlanden b. Stuttgart, W.E. Weinmann Druckerei GmbH (1972).

5.22 Ericsson, U.; Schneider, W.; Vogel, K.: The problem of privacy in a computer based integrated health care information system. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDFINO 74, Preprints, 649–652.

5.23 Evans, A. jr.; Kantrowitz, W.; Weiss, E.: A user authentication scheme not requiring secrecy in the computer. Commun. ACM 17 (1974), 437–442.

5.24 Fano, R.M.: On the social role of computer communications. Proc. IEEE 60 (1972), 1249–1253.

5.25 Farr, M.A.L.; Chadwick, B.; Wong, K.K.: Security for computer systems. Manchester, National Computing Centre (1973).

5.26 Feige, E.L.; Watts, H.W.: Protection of privacy through microaggregation. In: Bisco, R.L. (Ed.): Data bases, computers, and the social sciences. New York, J. Wiley (1970), 261–272.

5.27 Feige, E.L.; Watts, H.W.: An investigation of the consequences of partial aggregation of micro-economic data. Econometrica 2 (1972), 343–360.

5.28 Feistel, H.: Cryptography and computer privacy. Sci. Amer. 228,5 (1973), 15–23.

5.29 Fellegi, I.P.: On the question of statistical confidentiality. J. Amer. Statist. Ass. 67 (1972), 7–18.

5.30 Fellegi, I.P.; Phillips, J.L.: Statistical confidentiality: some theory and applications to data dissemination. Ann. Econ. Soc. Measurement 3 (1974), 399–409.

5.31 Fischer, Th.; Helmbock, J.M.: Data privacy and data security in Kiel KIS. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDINFO 74, Preprints, 197 199.

5.32 Friedman, T.D.: The authorization problem in shared files. In: Hoffman, L.J. (Ed.): Security and privacy in computer systems. Los Angeles, Melville Publ. Co. (1973), 159–185.

5.33 Friedman, T.D.: The authorization problem in shared files. IBM Syst. J. 9 (1970), 258–280.

5.34 Hansen, M.H.: Insuring confidentiality of individual records in data storage and retrieval for statistical purposes. AFIPS Conf. Proc. 39 (1971), 579–585.

5.35 Hanson, C.: Privacy and computerized health information. Med. Care 13 (1975), 273–274.

5.36 Helmbock, J.M.: Datenschutz und Datensicherung (Data privacy and data security) unter TP-Bedingungen. Fachtagung der Gesellschaft für Informatik und Gesellschaft für Med. Dokumentation und Statistik, 12.–14.10.1972, Med. Hochsch. Hannover (mimeographed paper).

5.37 Hentschel, B.; Gliss, H.; Bayer, R.; Dierstein, R.: Datenschutzfibel unter besonderer Berücksichtigung des Personalwesens. Köln, Bachem (1974).

5.38 Hoffman, L.J.; Miller, W.F.: Getting a personal dossier from a statistical data bank. Datamation 16 (1970) 5, 74–75.

5.39 Hoffman, L.J.: IBM's resource security system (RSS). In: Hoffman, L.J. (Ed.): Security and privacy in computer systems. Los Angeles, Melville Publ. Co. (1973), 379–392.

5.40 Hoffman, L.J.: The formulary model for flexible privacy in computer systems. Los Angeles, Melville Publ. Co. (1973), 168–215.

5.41 Jacobs, G.: Die Unwirksamkeit der Anonymisierung von Individualdaten – dargestellt am Beispiel der Amtlichen Studentenstatistik. Öffentl. Verwaltung u. Datenverarbeitung 3 (1973), 258–261.

5.42 Jourard, S.M.: Self-disclosure. An experimental analysis of the transparent self. New York, Wiley-Interscience (1971).

5.43 Karhausen, M.O.; Müller, P.J.: Datenbank, Datentransparenz und Datenschutz. Nachr. Dok. 23 (1972), 148–153.

5.44 Liu, P.T.; Chow, L.P.; Mosley, W.H.: Use of the randomized response technique with a new randomizing device. J. Amer. Statist. Ass. 70 (1975), 329–332.

5.45 Molho, L.M.: Hardware aspects of secure computing. In: Hoffman, L.J. (Ed.): Security and privacy in computer systems. Los Angeles, Melville Publ. Co. (1973), 351–365.

5.46 Müller, P.J.: Datenschutz und Sicherung der Individualdaten der empirischen Sozialforschung. Datenverarbeitung in Steuer, Wirtschaft und Recht 3 (1974), 2–11.

5.47 Nargundkar, M.S.; Saveland, W.: Random-rounding: a means of preventing disclosure of information about individual respondents in aggregate data. Ottawa, Statistics Canada (1972).

5.48 Nargundkar, M.S.; Saveland, W.: Random-rounding to prevent statistical disclosures. Amer. Statist. Ass. Proc., Soc. Statist. Section (1972), 382–385.

5.49 Palme, J.: Software security. Datamation 20 (1974), 51–55.

5.50 Peters, B.: Security considerations in a multi-programmed computer system. AFIPS conference proceedings, 1967 spring joint computer conference, vol. 30. Washington, Thompson Books (1967), 283–286.

5.51 Petersen, H.E.; Turn, R.: System implications of information privacy. In: Hoffman, L.J. (Ed.): Security and privacy in computer system. Los Angeles, Melville Publ. Co. (1973), 76–95.

5.52 Petersen, H.E.; Turn, R.: System implications of information privacy. AFIPS conference proceedings, 1967 spring joint computer conference, vol. 30. Washington, Thompson Books (1967), 291–300.

5.53 Reed, I.S.: Information theory and privacy in data banks. AFIPS Conf. Proc. 42 (1973), 581–587.

5.54 Schlörer, J.: Schnüffeltechniken und Schutzmaßnahmen bei statistischen Dantenbank-Informationssystemen mit Dialogauswertung. Materialien Nr. 29 der Abteilung für Medizinische Statistik, Dokumentation und Datenverarbeitung der Universität Ulm (1974).

5.55 Schlörer, J.: Identification and retrieval of personal records from a statistical data bank. Meth. Inform. Med. 14 (1975), 7–13.

5.56 Schlörer, J.: On statistical confidentiality: Minimum query language requirements for tracker building. Materialien Nr. 33 der Abteilung für Medizinische Statistik, Dokumentation und Datenverarbeitung der Universität Ulm (1975).

5.57 Schlörer, J.: Zum Statistikgeheimnis: Risiken und Schutz statistischer Datenbanken. Datenv. im Recht (DVR) 5 (1976), 203–248.

5.58 Schlörer, J.: On statistical confidentiality: a threat-monitoring scheme for on line dialogue. Meth. Inform. Med. 15 (1976), 36–42.

5.59 Schulze, J.H.: Datenschutz in der Datenverarbeitung. IBM Nachr. 21 (1971), 640–645.

5.60 Selbmann, H.-K.: Ein Datenbanksystem zur Auswertung statistischer Datenbestände. Materialien Nr. 15 der Abteilung für Medizinische Statistik, Dokumentation und Datenverarbeitung der Universität Ulm (1973).

5.61 Selbmann, H.-K.: Bitstring processing for statistical evaluation of large volumes of medical data. Meth. Inform. Med. 13 (1974), 61–64.

5.62 Shapiro, N.Z.: Privacy and security in databank systems: Measures of effectiveness, costs, and protector-intruder interactions. In: Hoffman, L.J. (Ed.): Security and privacy in computer systems. Los Angeles, Melville Publ. Co. (1973), 267–286.

5.63 Shepard, R.: Algebraic cryptography – A convenient technique for maintaining privacy of files. In: Fuchs, G.; Wagner, G. (Eds.): Krankenhaus-Informationssysteme. Stuttgart–New York, Schattauer (1972), 233–240.

5.64 Thome, R.: Protection and confidentiality of medical data. I. Efficient data protection through projectspecific combination of methods. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDINFO 74, Preprints, 189–191.

5.65 Turn, R.: Privacy and security in personal information databank systems. Santa Monica (Calif.), The Rand Corporation R-1044-NSF (1974).

5.66 Turn, R.; Shapiro, N.Z.: Privacy and security in databank systems – measures of effectiveness, costs, and protector-intruder interactions. AFIPS Conf. Proc. 41 I (1972), 435–444.

5.67 Van Tassel, D.: Cryptographic techniques for computers: substitution methods. Inform. Stor. Retr. 6 (1970), 241–249.

5.68 Wagner, G.: Das Problem der Vertraulichkeit im Zeitalter des Computers. In: Derbolowski, U.; Stephan, E. (Eds.): Die Wirklichkeit und das Böse. Hamburg, Christians (1970), 105–111.

5.69 Wagner, G.: Medizinische Datenbanken und ihre Problematik. In: Hollberg, N.; Pleuss, B.; Rittersbacher, H. (Eds.): Computer: Aufgaben im Gesundheitswesen. Berlin–Heidelberg–New York, Springer (1973), 22–31.

5.70 Ware, W.H.: Security and privacy in computer systems. AFIPS conference proceedings, 1967 spring joint computer conference, vol. 30. Washington, Thompson Books (1967), 279–282.

5.71 Ware, W.H.: Security and privacy: Simularities and differences. AFIPS conference proceedings, 1967 spring joint computer conference, vol. 30. Washington, Thompson Books (1967), 287–290.

5.72 Warner, S.L.: Randomized response: a survey technique for eliminating evasive answer bias. J. Amer. Statist. Ass. 60 (1965), 63–69.

5.73 Westin, A.F.: Privacy and freedom. New York, Atheneum (1970).

5.74 Yanez, L.M.: Introduction of a user-oriented THIS into a community hospital setting – confidentiality and security. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDINFO 74, Preprints, 201–206.

6. *Information Retrieval*

6.1 Abbel, R.: Information analysis centers. Nachr. Dok. 21 (1970), 95–97.

6.2 Back, H.B.: What information dissemination studies imply concerning the design of on-line reference retrieval system. J. Amer. Soc. Inform. Sci. 23 (1972), 156–163.

6.3 Barher, A.S.; Barraclough, F.D.; Gray, W.A.: On-line information retrieval as a scientists tool. Inform. Stor. Retr. 9 (1974), 429–440.

6.4 Batman, R.B.: End-user data control. Inform., p. 2 (New York), 3 (1974), 22–33.

6.5 Bernal, J.D.: The transmission of scientific information: a user's analysis. In: Proc. Intern. Conf. on Scientif. Inf. Washington (1959), 77–95.

6.6 Bock, M.; Hausen, H.-L.; Konrad, E.; Zuse, H.: FAKYR – An on-line information retrieval system. Conference on information sciences and systems, Baltimore (Maryland), 1975.

6.7 Bollmann, P.; Konrad, E.: Zur Anwendung von graphentheoretischen Cluster-Algorithmen auf das Begriffsnetz eines Wörterbuches. Conference "Algorithmen auf Graphen" of the German chapter of ACM (1975).

6.8 Borko, H.: Utilization of on-line interactive display. Santa Monica (Calif.), System Development Corp. ESP-2527, AD-640 652 (1966).

6.9 Borko, H.; Bunaugh, H.P.: Interactive displays for document retrieval. Inform. Display 3 (1966), 47–90.

6.10 Borman, L.; Mittman, B.: Interactive search of bibliographic data bases in adacemic environment. J. Amer. Coc. Inform. Sci. 23 (1972), 164–171.

6.11 Branscomb, L.M.: The misinformation explosion: Is the literature worth reviewing? Scientific Research 3 (1968), 49–52.

6.12 Brenner, E.H.: Some level of interaction in information science. Inform. Stor. Retr. 7 (1971), 211–215.

6.13 Buckley, J.D.; Gledhill, V.X.; Mathews, J.D.; Mackay, I.R.: SEARCH – A language for retrieval and statistical analysis of medical records. Comput. biomed. Res. 6 (1973), 235–244.

6.14 Chai, D.T.: An information retrieval system using keyword dialog. Inform. Stor. Retr. 9 (1973), 373–387.

6.15 Chang, S.K.: An interactive system for chinese character generation and retrieval. IEEE Trans. Syst. Man Cybern. 3 (1973), 257–265.

6.16 Cleverdon, C.W.; Keen, E.M.: Factors determining the performance of indexing systems. ASLIB-Cranfield Research Project Reports, Vols. 1 and 2, Cranfield (England), (1966).

6.17 CODATA: Second international CODATA conference on generation, collection, evaluation, and dissemination of numerical data for science and technology. Codata newsl. (Frankfurt), (1970), 2–7.

6.18 CODATA: Proceedings of the third international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology, Le Creusot, 1972. Frankfurt (1973).

6.19 CODATA: Proceedings of the 4th international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology, Tsakhcadzor, 1974. CODATA bull. (Paris), 14 (1975), 1–171.

6.20 CODATA Task Group, Paris: Guide for the presentation in the primary literature of numerical data derived from experiments. Report of the CODATA Task Group on publication data in the primary literature, September 1973. CODATA bull. (Frankfurt), 9 (1973), 1–6.

6.21 Codd, T.: A relational model of data for large shared data banks. Comm. ACM 13 (1970), 377–387.

6.22 Colby, K.M.; Enea, H.: Heuristic methods for computer understanding of natural language in context-restricted on-line dialogue. Math. Biosci. 1 (1967), 1–25.

6.23 Correll, W.D.; Evans, G.D.: COMPACT (Computerization of world facts). Menlo Park (Calif.), Stanford Research Institute (1965).

6.24 Curtice, R.M.; Jones, P.E.: An operational interactive retrieval system. Cambridge (Mass.), A.D. Little (1969).

6.25 Darby, R.L.: Information analysis centers as a source for information and data. Special Libraries 59 (1968), 91–97.

6.26 Drew, D.L.; Summit, R.K.; Tanaka, R.I.; Whiteley, R.B.: An on-line technical library reference retrieval system. Amer. Doc. 17 (1966), 3–7.

6.27 Egorova, N.A.: A functional predicate language for factographic data (translat. from Russ.). Autom. docum. and math. linguist. 9 (1975), 63–67.

6.28 Ferschl, F.: Probleme und Möglichkeiten einer statistischen Dokumentation. Metrika 14 (1969), 277–292.

6.29 Fraser, J.B.: On the role of natural language in man-machine communication. In: Walker, W.D. (Ed.): Information system science and technology. Washington, Thompson Books (1967), 21–28.

6.30 Freeman, R.R.; Atherton, P.: AUDACIOUS – An experiment with an on-line interactive reference retrieval system using the universal decimal classification as the index language in the field of nuclear science. New York, American Institute of Physics (1968).

6.31 Garvin, D.: The information analysis center and the library. Spec. libr. 62, (1971), 17–23.

6.32 Golashvili, T.: Statistical distribution of property data in published literature. In: CODATA: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 86–98.

6.33 Gothier, W.; Ecker, F.; Sellmair, P.: Implementation of an inquiry language for a medical data bank system. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDINFO 74, Preprints, 309–312.

6.34 Harm, K.; Höhne, K.H.; Dahlmann, K.: Computeranwendung im hämatologischen Laboratorium. II. Rückgewinnung der Meßwerte für die Erfordernisse der Routine und für die wissenschaftliche Auswertung. Ärztl. Lab. 19 (1973), 237–245.

6.35 Hercz, L.; Laszlo, Ch.A.; Reesal, M.R.: A computerized information system for pathology. Meth. Inform. Med. 14 (1975), 181–188.

6.36 Humphrey, S.M.: Searching the MEDLARS citation file on-line using ELHILL 2 and STAIRS: a comparison. Inform. Stor. Retr. 10 (1974), 321–329.

6.37 Jacobitz, K.; Bogenstätter, P.: Free text synthesis system (FTSS). Meth. Inform. Med. 13 (1974), 10–18.

6.38 Karpinski, B.H.S.; Bleich, H.L.: MISAR – A miniature information storage and retrieval system. Comput. biomed. Res. 4 (1971), 655–660.

6.39 Katzer, J.: The development of a semantic differential to assess user's attitudes towards an on-line interactive reference retrieval system. J. Amer. Soc. Inform. Sci. 23 (1972), 122–127.

6.40 Kljucarev, G.V.: Mnemonic information language for the data banks of automated factographic retrieval systems (translated from Russ.). Autom. docum. and math. linguist. 8 (1974), 10–18.

6.41 Kochen, M.: Principles of information retrieval. New York, J. Wiley (1974).

6.42 Koehler, C.O.; Schadewaldt, K.: AVAS – General variable evaluation system. In: Anderson, J.; Forsythe, J.M. (Eds.): MEDINFO 74, Preprints, 407–410.

6.43 Konrad, E.: Dynamische Dokumentenräume. Berlin, Heidelberg, New York, Springer = Lecture Notes in Economics and Mathematical Systems. Vol. 78 (1973), 499–502.

6.44 Konrad, E.: Deduktionsprozesse in Frage-Antwort-Systemen. In: Veenker, G. (Ed.): Künstliche-Intelligenz-Forschung in der BRD. Bonn (1975), 3–14.

6.45 Krieg, A.F.; Henry, J.B.; Stratakos, S.M.: Analysis of clinical pathology data by means of a user-oriented on-line data system. In: Enslein, K. (Ed.): Data acquisition and processing in biology and medicine. Oxford, Pergamon Press (1968), 163–172.

6.46 Lancaster, F.W.: Interaction between requesters and a large mechanized retrieval system. Inform. Stor. Retr. 4 (1968), 239–252.

6.47 Lancaster, F.W.; Rapport, R.L.; Kiffin, P.J.: Evaluating the effectiveness of an on-line natural language retrieval system. Inform. Stor. Retr. 8 (1972), 223–245.

6.48 Lancaster, F.W.: Information retrieval systems. New York, J. Wiley (1968).

6.49 Landau, R.: On-line interactive information systems. Proceedings of the 6th annual national colloquium of information retrieval 1969. Philadelphia, Medical Documentation Service, The College of Physicians of Philadelphia (1969), 359–372.

6.50 Langefors, B.: Theoretical analysis of information systems. Philadelphia, Auerbach Publ. (1973).

6.51 Martin, T.H.; Carlisle, J.; Treu, S.: The user interface for interactive bibliographic searching: an analysis of the attitudes of nineteen information scientists. J. Amer. Soc. Inform. Sci. 24 (1973), 142--147.

6.52 McCarn, D.B.; Leiter, J.: On-line services in medicine and beyond. Science 181 (1973), 318–324.

6.53 Mitchell, P.C.: The design of a on-line interactive document retrieval system and the general utility of the design techniques. Washington, Ph. D. Thesis, State University (1971).

6.54 Mitchell, P.C.; Rickman, J.T.; Walden, W.E.: SOLAR - A storage and online automatic retrieval system. J. Amer. Soc. Inform. Sci. 24 (1973), 347–358.

6.55 Negus, A.E.; Hall, J.L.: Towards an effective on-line reference retrieval system. Inform. Stor. Retr. 7 (1971), 249–270.

6.56 Olson, P.E.: Mechanisation of library procedures in the mechanized medical library: XV. A study of the interaction on non-librarian searchers with the MEDLINE retrieval system. Bull. med. Libr. Ass. 63 (1975), 35–41.

6.57 Oren, S.S.: A mathematical theory of man-machine text editing. IEEE Trans. Syst. Man Cybern. 4 (1974), 258–267.

6.58 Paisley, W.: Improving a field-based "eric-like" information system. J. Amer. Soc. Inform. Sci. 22 (1971), 399–408.

6.59 Popov, I.I.; Romanenko, A.G.; Sumarokov, L.N.: A model and algorithm for two-level control of information flows in an automated information system servicing a branch of industry (translated from Russ.). Autom. docum. and math. linguist. 8 (1974), 26–39.

6.60 Salton, G.: A document retrieval system for man-machine interaction. New York, Proceedings of the ACM, (1964) pages L 2.3-1 to L 2.3-20.

6.61 Salton, G.: Evaluation problems in interactive information retrieval. Inform. Stor. Retr. 6 (1970), 29–44.

6.62 Salton, G.: Automatic information organization and retrieval. New York, McGraw-Hill (1968).

6.63 Salton, G. (Ed.): The SMART retrieval system. Prentice Hall (1971).

6.64 Saxl, J.: An innovative medical information storage and retrieval system. Hosp. Progr. 52 (1971), 44–48.

6.65 Schneider, H.J.: Die Strukturiertheitshierarchie bei den verschiedenen Modellen einer Problemlösung, Workshop "Problemgegebene Datenstrukturen". In: Durchholz; Konrad; Richter; Schneider (Eds.): Birlinghoven (1973), 4–18, 156–160.

6.66 Schneider, H.J.: Bibliotheksautomatisierung und Dokumentationssysteme aus der Sicht der Informatik. In: Deutscher Dokumentartag 1973. München, Verlag Dokumentation (1974), 74–91.

6.67 Schultz, L.: Breaking the communication barrier between searcher and literature file: An interactive guide. J. Amer. Soc. Inform. Sci. 25 (1974), 3–9.

6.68 Selbmann, H.-K.: Bitstring processing for statistical evaluation of large volumes of medical data. Meth. Inform. Med. 13 (1974), 61–64.

6.69 Senko, M.E.; Altman, E.B.; Astrahan, M.M.; Fehder, P.L.: Data structures and accessing in data-base systems. IBM Systems J. 12 (1973), 30–93.

6.70 Skinner, C.W.: A heuristic approach to inductive inference in fact retrieval systems. Comm. ACM 17 (1974), 707–712.

6.71 Soben, P.; Tidball, C.S.: MEDLEARN: An orientation to MEDLINE. Bull. med. Libr. Ass. 62 (1974), 92–94.

6.72 Summit, R.K.: DIALOG II – User's manual. Palo Alto, Lockheed Missiles and Space Co., Report 6-77-67-14 (1967).

6.73 Summit, R.K.: DIALOG – An operational on-line reference retrieval system. Proceedings of 22nd national conference, association for computing machinery. Washington, Thompson Books (1967), 51–56.

6.74 Summit, R.K.: ERIC on-line retrieval system. Use of the DIALOG on-line information retrieval system with ERIC research in education files. Palo Alto, Lockheed Research Lab., N-6R-70-1 (1970).

6.75 Sundgren, B.: The infological approach to data bases. In: Falkenberg; Klutentreter (Eds.): Datenmodelle und Systementwürfe für Datenbanksysteme. Birlinghoven, Summer School (1974), 133–174.

6.76 Tague, J.M.: A bayesian approach to interactive retrieval. Inform. Stor. Retr. 9 (1973), 129–142.

6.77 Vallee, J.F.: DIRAC – An interactive retrieval language with computational interface. Inform. Stor. Retr. 6 (1970), 387–399.

6.78 Wall, R.A.: Indexing language structure for automated retrieval. Inform. Stor. Retr. 9 (1973), 607–617.

6.79 Weizenbaum, J.: ELIZA – A computer program for the study of natural language communication between man and machine commun. ACM 9 (1966), 36–45.

6.80 Williams, J.H., jr.: Functions of a man-machine interactive information retrieval system. J. Amer. Soc. Inform. Sci. 22 (1971), 311–317.

7. *Information Analysis, Evaluation of Data*

7.1 Golashvili, T.V.: Problems of compilation and evaluation of property data of substances and materials. 2. Internat. CODATA conf. on generation, coll., eval. and dissemination of numerical data for sci. and technol., St. Andrews, 1970, CODATA. J. of the ASIS 24 (1973), 57–64.

7.2 Goundsmit, S.A.: Is the literature worth retrieving? Physics Today 19 (1966), 52–55.

7.3 Kiefer, L.J.: The information analysis and the creation of reliable data. In: Proceedings of the Office of Critical Tables Advisory Board Meeting, 1968. Washington, National Academy of Sciences (1969), 133–140.

7.4 Klemm, W.: CODATA: Internationale Koordination naturwissenschaftlicher und technischer Datenzentren. Umschau (1968), 404.

7.5 London, G.: The publication inflation. American Documentation 19 (1966), 137–141.

7.6 May, D.H.: The role of the office for scientific and technical information (OSTI). Aslib Proc. 21 (1969), 397–399.

7.7 Meyer, H.; Müller, R.; Hartmann, K.: Organisationsprojekt eines Faktenrecherchensystems für die Prognose. Rechentechnik Datenverarbeitung, Beih. 2 (1969), 27–32.

7.8 Office of Critical Tables. NAS-NRC: Continuing numerical data projects. A survey and analysis. Washington, National Academy of Sciences (1966).

7.9 Rossini, F.D.: The ICSU committee on data for science and technology (CODATA). J. Chem. Doc. 10 (1970), 261–264.

7.9a Sandor, L., Wagner, G., Wolff-Terroine, M.: Versuch einer Strukturanalyse der Krebsliteratur auf der Basis von SABIR-C. Z. Krebsforsch. 80 (1973), 69–82.

7.10 Schäfer, C.: Die Bedeutung der Datendokumentation und Datenbewertung für das Informationsbankensystem der Bundesrepublik Deutschland. In: Das Informationsbankensystem. Bd. 2. Materialband. Köln, Heymanns (1971), 95–108.

7.11 Schäfer, C.: Datendokumentation und Datenbewertung. In: Dokumentation und Information. Frankfurt, Umschau-Verlag (1971), 153–160.

7.12 Schäfer, C.: The role of information analysis centres in science and technology. In: CODATA, central office: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 75.

7.13 Schäfer, C.; Larink, J.: Joint WFEO/CODATA study on engineering data. In: CODATA: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 49–50.

7.14 Science Communication, Inc.: Preliminary census of scientific and technical data activities, parts A, B, C. Springfield (Va.), CFSTI, AD 670 (1968), 607.

7.15 Simpson, G.S., jr.: Information analysis center viewpoint on wordage problems: amount, languages, and access. J. Chem. Doc. 12 (1972), 97–98.

7.16 Soergel, D.; Furmaniak, K.: The description of statistical tables: A problem in data documentation. In: North, J.B. (Ed.): The information conscious society. 33rd annual meeting Philadelphia, 1970. Washington, Amer. Soc. for Inform. Sci. (1970), 331–336 = Proc. of the Amer. Soc. for Inform. Sci., 7.

7.17 Tongeren, van: The role of the publishing industry in the production and dissemination of critical reviews and data compilations. In: CODATA: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 96–97.

7.18 Waddington, G.: A world system of evaluated numerical data for science and technology. J. Chem. Doc. 7 (1967), 20–23.

7.19 Waldo, W.H.: Searchers want facts not folios – retrieve data not documents – the needle is dull – sharpen it with automation. In: Automation and scientific communication. ADI Ann. Meeting 1963. Short papers part 2. Washington, ADI (1963), 207–209.

7.20 Weinberg, A.M.: Second thoughts on scientific information. College and Research Libraries 25 (1964), 463–471.

7.21 Weizman, H.M.: Needs of American Chemical Society members for property data. J. Chem. Doc. 7 (1967), 9–14.

7.22 Wilhelmi, O., jr.; Brown, L.: The information analysis center: Key to better use of the information resource. J. Chem. Doc. 8 (1968), 106–109.

7.23 Wolfe, W.N. et al.: How can effectiveness of analysis centers be measured?, panel discussion, proceedings of the first ad hoc forum of scientific and technical IAC managers, directors, and professional analysts, Columbus, Ohio 1965. Springfield (Va.), NTIS (1966), CONF 651131.

8. *Presentation of existing Data Banks*

Administration

8.1 Gfeller, A.: Datenbanken in öffentlichen Verwaltungen. Neue Zürcher Ztg., Zürich 193 (1972), 163 (16.6.), 81.

8.2 IBM (Ed.): Informationssysteme und Datenbanken. Eine Auswahl von Vorträgen aus Seminaren für die öffentliche Verwaltung. Bonn, IBM (1971).

8.3 Thomas, U.: Datenbanken in der öffentlichen Verwaltung. Bonn, Asgard-Verl. (1974) = OECD-Informatik Studien 1.

Law and Police

8.4 Arbeitsgruppe "EDV und Recht": Die Informationsmacher. Zur Fragwürdigkeit einer privaten Datenbank für Rechtsdokumentation. Berlin, Juradat (1970).

8.5 Conradi, J.: Die elektronische Steuerrechtsdatenbank der DATEV. Datenverarb. in Steuer, Wirtsch. u. Recht 4 (1975), 98–109.

178

8.6 Datenverarbeitungsorganisation der Steuerbevollmächtigten für die Angehörigen des steuerberatenden Berufs in der Bundesrepublik Deutschland (DATEV); IBM: Ein Beitrag zur Bewältigung der Informationskrise. In: Recht. 49. Deutscher Juristentag, Düsseldorf, 1972. Nürnberg (1972).

8.7 Puka, K.: Jahresabschluß und Computer im DATEV-System. Teil 3: Individuelles Blitzprogramm-Kennziffernteil. Datenverarb. in Steuer, Wirtsch. u. Recht 4 (1975), 237–243.

8.8 Seifert, P.: Polizeiliches Informationssystem (POLIS). In: IBM: Informationssysteme und Datenbanken. Eine Auswahl von Vorträgen aus Seminaren für die Öffentliche Verwaltung. Bonn, IBM (1971), 229–238.

8.9 Simitis, S.: Informationskrise des Rechts und Datenverarbeitung. Karlsruhe, Müller (1970) = Recht – Justiz – Zeitgeschehen (RJZ), 7 (1970).

Social

8.10 Eisner, M.: A researcher's overview of the TROLL/1-system. Cambridge (Mass.), (1972).

8.11 Küppers, R.: Warum Sozialrechtsdatenbank? Bundesarb.bl. 23 (1972), 603–608.

8.12 Siemens, Bereich Datenverarbeitung, München: Sozialrechtsdatenbank bei Bundessozialgericht mit dem Informationssystem GOLEM. München, Siemens (1973).

8.13 Winkler, P.: Die Sozialdatenbank. In: Krückeberg, F.; Walder, W.; Brandt, B.A. (Eds.): Computer und Gesellschaft. Stuttgart, Frankfurt (1974), 105.

8.14 Zschau, H.: Sozialdatenbank – Ein Instrument der Sozialplanung. In: Data Report 9 (1974), 3, 4.

Traffic

8.15 Wilkenloh, F.: Planungsarbeiten für ein Verkehrsinformationssystem. Der Aufbau einer Dokumentations- und Verkehrsdatenbank als Hauptbestandteile eines Informations- und Entscheidungssystems im Verkehrswesen. Data Rep. 7 (1972), 3, 24–29.

Chemistry

8.16 Ash, J.E.; Hyde, E.: Chemical information systems. Chichester, Horwood (1975).

8.17 Bonnier, E.: Operation of a thermochemical data bank. In: CODATA: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 75–78.

8.18 Clamroth, R.; Fendel, H.; Hofmann, W.: Dokumentation von Werkstoffprüfdaten auf dem Kautschukgebiet unter Einsatz von elektronischen Datenverarbeitungsanlagen. Nachr. f. Dok., 23 (1972), 119–123.

8.19 Dubois, J.E.; Laurent, D.; Aranda, A.: DARC-System – XVI-Theory of topology-information. 1. Pertubation of ordered and concentrical limited environments method. 2. Pelco method-procedure for establishing correlations of topology-information. J. Chim. Phys. 70 (1973), 1608–1616.

8.20 Dubois, J.E.; Laurent, D.; Viellard, H.: DARC-System-description by concentric propagation from a limited environment and modular and nulinear method of representing a symmetrical graph. Paris, C.R. Acad. Sciences, A. 270 (1970), 228.

8.21 Dubois, J.E.: DARC system in chemistry. In: Wipke, W.T. (Ed.): Computer representation and manipulation of chemical information. New York, J. Wiley (1974).

8.22 Dubois, J.E.: French national policy for chemical information and DARC-System as a potential tool of this policy. J. Chem. Doc. 13 (1973), 8.

8.23 Dubois, J.E.; Laurent, D.; Fatome, M.; Andrieu, L.; Pignalosa, C.: Automatisme. Août (1975).

8.24 Geiss, F.; Bourdeau, P.: ECDIN, an EC data bank for environmental chemicals. Ispra, European Joint Res. Cent. (1974).

8.25 Kennard, O.; Allen, F.H.; Motherwell, W.D.S.; Town, W.G.; Watson, D.G.: Cambridge Crystallographic Data Centre. 2. Structural data file. J. Chem. Doc. 13 (1973), 119–123.

8.26 Koenig, M.E.D.: On-line substructure searching using fragment code – a proposal. J. Chem. Doc. 14 (1974), 127–130.

8.27 Koptjug, V.A.; Drobyshev, Yu.P.: Computerized molecular spectroscopy data handling in the Novosibirsk Scientific Centre. In: CODATA: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 67–70.

8.28 Meyer, E.F.: Storage and retrieval of macromolecular structural data. Biopolymers 13 (1974), 419–422.

8.29 Schliessinger, B.S.: Physicochemical property data treatment. J. Chem. Doc. 9 (1969), 20–24.

8.30 Sulov, L.M.; Ermakova, I.A.: Extraction of data on properties of compounds using integrated processing of literature on synthetic organic chemistry. Sci. and tech. inform. process (1974), 69–74.

Engineering

8.31 Born, A.: The iron-steel data base as an example of a data base scheme. Paper presented on the symposium on: man-machine communication for scientific data handling. Organized by the task group on computer use of the ICSU committee on data for science and technology (CODATA). Freiburg, Juli 22–27, 1973.

8.32 Carlowitz, B.: Datenbank für Kunststoffeigenschaften. Kunststoffe 64 (1974), 493–496.

8.33 Chapman, R.B.: Impact of improved engineering data and information systems on engineering practices. In: Speight, E.Y. (Ed.): Proceedings of the national engineering information conference, 1969. New York, Engineers Joint Council (1970), 16–19.

8.34 Dathe, G.; Dreckmann, K.-H.: Entwurf eines Datenbanksystems für normierte Kennwerte von Eisen- und Stahlwerkstoffen. Berlin–Heidelberg–New York, Springer, Lecture Notes in Computer Science 26 (1975), 484–493.

8.35 Jurk, H.; Mec, N.; Preuß, E.: Das Informationssystem für Werkstoffkennwerte und ökonomischen Materialersatz. Die Wirtschaft, 28 (1973) 40, 12–13.

8.36 Longenbaugh, R.A.; MacMillin, N.E.: A data manager looks at the development of the Colorado Water Data Bank. Inform., p. 2, 3 (1974), 34–47.

8.37 Mommertz, K.H.; Radestock, J.; Kremer, K.J.: Entwicklung einer Datenbank als Informationssystem über die Eigenschaften von Stählen. Stahl u. Eisen, 90 (1970), 1339–1346.

8.38 Nemoda, L.: Aspects of the organization and operation of the central data and information bank of the National Institute for Scientific and Technical Information and Documentation (INID). Probl. de informare si documentare, Bukarest 7 (1973), 730–755.

8.39 Shilito, B.J.: Department of defense engineering data programs and systems. In: Proc. Nat. Engineer. Inform. Conf., 1969. New York (1970), 37–40.

8.40 Sturza, D.: The fact information system of the VEB Kombinat Robotron. Probl. de informare si documentare, Bukarest 7 (1973), 478–491.

8.41 Univ. of Southampton; Dep. of mach.eng.: International symposium information systems for designers. The proceedings of the meeting held at the University, 1971. Southampton (1971).

Geo Sciences

8.42 Ehrismann, W.; Kupfernagel H.: Aufbau eines integrierten Informationssystems der rohstoffbezogenen Meeresforschung. München, Zentralstelle

f. Luft- u. Raumfahrtdokum. u. -inform. (ZLDI), (1973) = Bundesministerium für Forschung und Technologie, FB M 73-04.

8.43 Jung, J.: The Strasbourg Stellar Data Centre. In: CODATA: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 42–44.

8.44 London, T.V.: Elektronische Datenverarbeitung in der Geologie. Endesvour 29 (1970), 125–128.

8.45 National Cartographic Information Center. Sci. inform. notes 5 (1973), 10–12.

8.46 Waranius, F.B.; Heiken, J.H.: Lunar Data Information Center. A shortcut to the riddle of the moon. Spec. libr. (New York) 66 (1975), 407–410.

8.47 Wilkins, G.A.: Numerical data for astronomers. In: CODATA: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 40–42.

Medicine and Biomedicine

8.48 Abrahamsson, S.; Berstroem, S.; Larsson, K.; Tillman, S.: Danderyd hospital computer system. 2. Total regional system for medical care. Comput. biomed. Res. 3 (1970), 30–46.

8.49 Ball, M.J.: An overview of total medical information systems. Meth. Inform. Med. 10 (1971), 73–82.

8.50 Barnett, G.O.: The modular hospital information system. In: Stacy, R.W.; Waxman, B.D. (Eds.): Computers in biomedical research, vol. 4. New York, Academic Press (1974).

8.51 Barnett, G.O.; Greenes, R.A.: Interface aspects of a hospital information system. Ann. N.Y. Acad. Sci. 161 (1969), 756–768.

8.52 Barnett, G.O.; Greenes, R.A.; Grossman, J.H.: Computer processing of medical text information. Meth. Inform. Med. 8 (1969), 177–182.

8.53 Bekey, A.; Schwarzt, M.D. (Eds.): Hospital information systems. New York, Marcel Dekker Inc. (1972).

8.54 Boehme, G.H.; Reichertz, P.; Wolters, E.: IRSD – Individual retrieval system for diagnoses. System-Design und Anwendung auf dem Gebiet der Endokrinologie. In: Fuchs, G.; Wagner, G. (Eds.): Krankenhaus-Informationssysteme. Stuttgart–New York, Schattauer (1972), 305–310.

8.55 Budd, P.J.: Planning a comprehensive hospital information system. Proceedings on automated data processing in hospitals. Conf. in Elsionore (Denm.), 1966. Stockholm (1966), 164–179.

8.56 Collen, M.: Hospital computer systems. New York, J. Wiley (1974).

8.57 Ehrengruber, H.; Spielmann, J.: Datenverarbeitung am Inselspital Bern. Med. Hyg. 31 (1973), 825–828.

8.58 Fink, H.: Computer-Hilfen in der Arzneimittelforschung. Münch. med. Wschr. 112 (1970), 1525–1532.

8.59 Frenzel, D.R.: Basisdokumentation diagnostischer und therapeutischer Patientendaten. Angew. Informatik 16 (1974), 122–128.

8.60 Fuchs, G.; Wagner, G. (Eds.): Krankenhaus-Informationssysteme – Erstrebtes und Erreichtes. Stuttgart–New York, Schattauer (1972).

8.61 Fuchs, G.: Probleme der Dokumentation medizinischer Patientendaten. In: Kschenka, W. u.a. (Eds.): Information und Dokumentation im Aufbruch. Pullach, Verl. Dokum. (1975), 76–90.

8.62 Giere, W.: Das Informationssystem der Deutschen Klinik für Diagnostik. Münch. med. Wschr. 117 (1975), 875–878.

8.63 Graul; Habermehl, A. (Eds.): Computersysteme in der Medizin. Systems in medicine. Lövenich/Krs. Köln, Dtsch. Ärzte-Verlag (1973).

8.64 Greenes, R.A.; Pappalardo, A.N.; Marble, C.W.; Barnett, G.O.: Design and implementation of a clinical data management system. Comput. biomed. Res. 2 (1969(, 469–485.

8.65 Greiller, R.; Pribilla, P.: EDV im Klinikum Großhadern. Data Report 9 (1974), 1, 15–20.

8.66 Griesser, G.: Das Kieler Klinik-Informations-System – Aufbau, Funktion und bisherige Erfahrungen. Computerunterstützte Informationssysteme für Krankenhäuser. IBM Seminare, Bad Liebenzell (1973) und Bad Godesberg (1972). Bonn, IBM (1973), 21–31.

8.67 Händler, H.; Harris, L.E.: Data documentation in the field of feed analysis and the international cooperation it includes. Quart., Bull of the Internat. Assoc. of Agric. Librarians and Documentalists, Bennkom. 18 (1973), 87–94.

8.68 Köhler, C.C.: Integriertes Krankenhaus-Informationssystem. Meisenheim/Glan, A. Hain (1973).

8.69 Lange, H.J.: Erfahrungen beim klinischen Einsatz des Informationssystems ISIS. Data Report 9 (1974), 4, 10–14.

8.70 Lehr, J.L.; Lodwick, G.S.; Nicholson, B.F.; Birznieks, F.B.: Experience with MARS (Missouri Automated Radiology System). Radiology 106 (1973), 289–294.

8.71 Mellner, C.; Selander, H.; Wolodarski, J.: The Karolinska Hospital information system. Meth. Inform. Med. 13 (1974), 125–140.

8.72 Perring, F.: The biological records centre – a data centre. Biol. J. of the Linnean Soc. (London) 3 (1971), 237–243.

8.73 Pirtkien, R.: Aufbau einer Datenbank für Auskünfte bei Vergiftungen. Informationssysteme in der Medizin. IBM Seminar, Bad Liebenzell (1971), 77–103.

8.74 Sauter, K.; Reichertz, P.L.; Zowe, W.: Die zentrale Patienten-Datenbank in einem integrierten Hospital-Informationssystem. Meth. Inform. Med. 11 (1972), 91–96.

8.75 Wagner, G.: Medizinische Datenbanken ind ihre Problematik. In: Hollberg, N.; Pleuss, B.; Rittersbacher, H. (Eds.): Computer: Aufgaben im Gesundheitswesen. Berlin–Heidelberg–New York, Springer (1973), 22–31.

8.75a Wagner, G.; Wolff-Terroine, M.: Das Krebsliteratur-Dokumentationssystem SABIR. Z. Krebsforsch. 78 (1972), 4–11.

8.75b Wolff-Terroine, M.: SABIR: Un systéme automatique de bibliographie, d'information et de recherche en carcinologie. Bull. Bibl. France 15 (1970), 169–176.

Standards

8.76 Abbel, R.: Das amerikanische National Standard Reference Data System. Ein Beispiel für die nationale Organisation der Dokumentation physikalischer und chemischer Daten. Nachr. f. Dokum. 25 (1974), 105–108.

8.77 Abbel, R.: Bericht über das amerikanische National Standard Reference Data System (NSRDE). Mit Kurzbeschreibungen einiger ausgewählter Dantenzentren. Frankfurt (1973).

8.78 Alt, F.L.: Information handling in the National Standard Reference Data System. Washington, U.S. Governm. Printing Office (1966) = NBS Techn. Note 290.

8.79 Bourcier, S.: On the interface between sciences and documentation in the plan for the international tables of selected constants. In: CODATA: Proceedings of the 3rd international CODATA conference on generation, compilation, evaluation and dissemination of data for science and technology. Le Creusot, 1972. Frankfurt (1973), 98–99.

8.80 Brady, E.L.: The National Standard Reference Data System. J. Chem. Doc. 7 (1967), 6–9.

8.81 Brady, E.L.; Wallenstein, M.B.: National Standard Reference Data System. Plan of operation. Washington, National Bureau of Standards (1964) = NSRDS-NBS 1.

8.82 Brady, E.L. (Ed.): Status report. National Standard Reference Data System. April 1966. Washington, U.S. Governm. Printing Office (1966) = NBS Techn. Note 289.

8.83 Deutsches Komitee für naturwissenschaftliche und technische Daten. Pressedienst BMW (München), (1967), 74.

8.84 National Bureau of Standards: National Standard Reference Data Program. Background information. Washington, NBS (1963) = Technical note 194.

8.85 Palme, W.: Methodische Aspekte und Erfahrungen bei der Erarbeitung synthetischer Kennziffern-Information. Informatik 18 (1971), 37–39.

8.86 Rossini, F.D.: Historical background of data compiling activities. J. Chem. Doc. 7 (1967), 2–6.

8.87 Weisman, H.M.: Information center profile office of standard reference data. Scientific Information Notes 2 (1970), 187–194.

Miscellaneous

8.88 Levitt, J.H.; Labarre, C.E.: Building a data file from historical archives. Comput. and the humanities 9 (1975), 77–82.

8.89 Lube, M.: Literaturdokumentation und Datendokumentation an der Universität Graz. Bericht über eine von November 1973 bis Jänner 1974 durchgeführte Erhebung. Graz (1974).

8.90 Recla, J.: Sport-Dokumentation im Durchbruch. Analysen und Aspekte der Literaturdokumentation, Datendokumentation und der Dokumentation audio-visueller Techniken. 3. internationaler Kongreß für Dokumentation der Sportwissenschaften 1970. Schielleiten-Graz. Bad Honnef, Osang (1971).

8.91 Rowe, J.S.; Ryan, T.: Library service from numerical data bases: The 1970 census as a paradig. College and res. libr. 35 (1974), 7–15.

List of Authors

Andersen, Dr. Per Klitgaard
Kollemosevej 2a, DK-2840 Holte

Appleyard, Raymond K.
Director General for Scientific and Technical Information and Information Management
Commission of the European Communities
Luxemburg

Bonnet, Jean-Claude
Laboratoire de Chimie Organique Physique, Faculté de Sciences
1, Rue Guy de la Brosse, F-75005 Paris

Dathe, Dr.-Ing. Gerd
Betriebsforschungsinstitut VDEh
Institut für angewandte Forschung GmbH
Sohnstr. 65, D-4000 Düsseldorf 1

Dreckmann, Dipl.-Ing. Karl-Heinz
Betriebsforschungsinstitut VDEh
Institut für angewandte Forschung GmbH
Sohnstr. 65, D-4000 Düsseldorf 1

Dubois, Jacques-Emile
Laboratoire de Chimie Organique Physique, Faculté de Sciences
1, Rue Guy de la Brosse, F-75005 Paris

Gaus, Wiss. Rat und Prof. Dr. Wilhelm
Universität Ulm
Schule für Medizinische Dokumentationsassistenten
Schloßbau 38, D-7900 Ulm-Wiblingen

Guaydier, Jean-François
Laboratoire de Chimie Organique Physique, Faculté de Sciences
1, Rue Guy de la Brosse, F-75005 Paris

Henzler, Dipl.-Math. Rolf
Deutsches Krebsforschungszentrum
Im Neuenheimer Feld 280, D-6900 Heidelberg 1

Killian, Dipl.-Ing. Dr. Klaus-Peter
Institut für Klinische Chemie im Klinikum Großhadern
Marchionistr. 15, D-8000 München 70

Konrad, Dr. Erhard
Technische Universität Berlin, Fachbereich Kybernetik
Straße des 17. Juni 135, D-1000 Berlin 12

Lockemann, Prof. Dr. Peter C.
Institut für Informatik II der Universität Karlsruhe
Zirkel 2, D-7500 Karlsruhe 1

Lutz, Dipl.-Math. Theo
Eichenweg 18, D-7300 Esslingen
and
IBM-Hauptverwaltung, D-7032 Sindelfingen

Schäfer, Dr. Christoph
Informationszentrum für Biologie
Forschungsinstitut Senckenberg
Senckenberg-Anlage 25, D-6000 Frankfurt (Main)

Schlörer, Dr. Jan
Universität Ulm, Klinische Dokumentation
Schloßbau 38, 7900 Ulm-Wiblingen

Schmidt, Reg.-Dir. Dr. Herbert
Bundesministerium für Arbeit und Sozialordnung
Rochungsstr. 1, D-5300 Bonn-Duisdorf

Weingarten, Dipl.-Ing. Werner
Medizinische Hochschule Hannover
Department für Biometrie und Medizinische Informatik
Karl-Wichert-Allee 9, D-3000 Hannover-Kleefeld

Wersig, Prof. Dr. Gernot
Institut für Medizinische Statistik und Dokumentation im Klinikum
Steglitz der Freien Universität Berlin
Hindenburgdamm 30, D-1000 Berlin 45